
Psicometria

PID_00241073

Julio Meneses
Maite Barrios
Albert Bonillo
Antoni Cosculluela
Luis Manuel Lozano
Jaume Turbany
Sergi Valero

**Julio Meneses**

Professor dels Estudis de Psicologia i Ciències de l'Educació de la Universitat Oberta de Catalunya (UOC) i investigador de l'Internet Interdisciplinary Institute (IN3). Llicenciat en Psicologia per la Universitat d'Oviedo, es va doctorar en el programa de Societat de la Informació i el Coneixement de la UOC i, actualment, treballa com a responsable de la metodologia en diversos projectes d'investigació a les ciències socials. La seva recerca s'orienta a l'aplicació dels mètodes de recerca quantitativa i a l'ús de tècniques d'anàlisi multivariant, particularment en les àrees de la introducció de les TIC als sistemes educatius, la relació dels infants i els joves amb la tecnologia i l'abandonament i el reingrés dels estudiants universitaris.

**Maite Barrios**

Professora agregada del Departament de Metodologia de les Ciències del Comportament de la Universitat de Barcelona (UB) i consultora dels Estudis de Psicologia i Ciències de l'Educació de la Universitat Oberta de Catalunya (UOC). Doctora en Psicologia per la UB. Desenvolupa la docència en l'àmbit de la metodologia impartint, entre d'altres, les assignatures de Psicomètria i Estadística aplicada a la psicologia i a informació i documentació. Coautora de diversos llibres i textos sobre estadística i tècniques de recerca. La seva recerca se centra, principalment, en la infometria, el biaix de gènere en ciència i els estudis bibliomètrics.

**Albert Bonillo**

Professor lector de l'Àrea de Metodologia de les Ciències del Comportament de la Facultat de Psicologia de la Universitat Autònoma de Barcelona (UAB) i consultor dels Estudis de Psicologia i Ciències de l'Educació de la Universitat Oberta de Catalunya (UOC). Doctor en Psicologia per la UAB, màster en Disseny i estadística en ciències de la salut. En la seva tesi doctoral proposa un conjunt d'algorismes per depurar dades ja registrades. La seva recerca s'orienta, entre altres àmbits, a la depuració de dades, la detecció de plagi acadèmic, i als factors explicatius i conseqüències dels dèficits en la funció executiva dels infants.

**Antoni Cosculluela**

Professor titular d'universitat del Departament de Metodologia de les Ciències del Comportament de la Facultat de Psicologia de la Universitat de Barcelona (UB) i consultor dels Estudis de Psicologia i Ciències de l'Educació de la Universitat Oberta de Catalunya (UOC). Doctor en Psicologia per la UB. La seva tasca docent s'ha desenvolupat en l'àmbit de la metodologia i els dissenys de recerca, i de l'anàlisi de dades en psicologia i en informació i documentació. Autor i coautor de diferents llibres d'estadística aplicada. Actualment la seva recerca s'orienta, entre altres camps, als estudis bibliomètrics en diferents àrees de coneixement.

**Luis Manuel Lozano**

Professor contractat doctor de l'Àrea de Metodologia de les Ciències del Comportament de la Universitat de Granada, membre del CIMCYC i consultor dels Estudis de Psicologia de la Universitat Oberta de Catalunya (UOC). Doctor en Psicologia per la Universitat d'Oviedo. Desenvolupa la docència en l'àrea de metodologia, i concretament imparteix les assignatures de Descripció i exploració de dades en psicologia i Tècniques d'anàlisi en recerques psicològiques. Els seus camps de recerca estan relacionats amb l'anàlisi freqüentista i bayesià de dades, l'elaboració de qüestionaris i models de teoria de la resposta als ítems.

**Jaume Turbany**

Professor titular d'universitat del Departament de Metodologia de les Ciències del Comportament de la Facultat de Psicologia de la Universitat de Barcelona (UB) i consultor dels Estudis de Psicologia i Ciències de l'Educació de la Universitat Oberta de Catalunya (UOC). Doctor en Psicologia per la UB. La seva tasca docent s'ha desenvolupat en l'àmbit de la metodologia i l'estadística aplicades a la psicologia i a la informació i documentació. Actualment la seva recerca s'orienta, entre altres camps, als estudis bibliomètrics en diferents àrees de coneixement, i a la recerca en innovació docent.

**Sergi Valero**

Psicòleg del Servei de Psiquiatria de l'Hospital Universitari Vall d'Hebron de Barcelona i consultor dels Estudis de Psicologia i Ciències de l'Educació de la UOC. Doctor en Psicologia per la Universitat Autònoma de Barcelona (UAB). La seva tasca

ca científica està emmarcada en la recerca bàsica i aplicada en l'àmbit de la psicologia, la psicopatologia i les demències. Les seves publicacions inclouen espais diversos de recerca, des de l'estudi de les dimensions de personalitat normal fins a la genètica, passant pels assajos clínics, la construcció i adaptació d'instruments de mesura psicològica, l'epidemiologia, les drogodependències o la neuropsicologia.

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats pel professor: Julio Meneses Naranjo (2016)

Segona edició: setembre 2016

© Julio Meneses, Maite Barrios, Albert Bonillo, Antoni Cosculluela, Luis Manuel Lozano, Jaume Turbany, Sergi Valero

Tots els drets reservats

© d'aquesta edició, FUOC, 2016

Av. Tibidabo, 39-43, 08035 Barcelona

Disseny: Manel Andreu

Realització editorial: Oberta UOC Publishing, SL

Dipòsit legal: B-15.978-2016

Es garanteix el permís per a copiar, distribuir i modificar aquest document segons els termes de la GNU Free Documentation License, Version 1.3 o qualsevol altra de posterior publicada per la Free Software Foundation, sense seccions invariants ni textos de la coberta anterior o posterior. Podeu consultar els termes de la llicència a <http://www.gnu.org/licenses/fdl-1.3.html>.

*“Mesura el que és mesurable i fes
mesurable el que no ho és.”*

Atribuït a Galileo Galilei

Introducció

Els tests formen part de la pràctica habitual dels professionals, docents i investigadors interessats per la mesura indirecta dels fenòmens psicològics. Al servei de l'avaluació psicològica, i en conjunció amb d'altres instruments com l'observació o l'entrevista, els tests tenen com a propòsit principal proporcionar les evidències necessàries que permetin als psicòlegs i a d'altres professionals vinculats amb les ciències socials i del comportament prendre decisions o orientar les seves intervencions en els diferents contextos en què duen a terme la seva activitat. Els tests compleixen, doncs, una funció important amb relació a l'avaluació i la intervenció psicològiques, i és per això que requereixen una atenció específica en els programes de formació dels futurs psicòlegs. Conèixer el seu funcionament, les seves propietats i les condicions en què els tests han de ser utilitzats de manera adequada i responsable són alguns dels reptes importants que abordarem en aquest text.

Per fer-ho, al llarg dels propers mòduls ens endinsarem en els aspectes teòrics i pràctics involucrats en la mesura indirecta dels fenòmens psicològics mitjançant tests que la psicometria ha proposat i sistematitzat en les darreres dècades. Com a branca de la psicologia, la psicometria és la disciplina científica encarregada del desenvolupament de teories, mètodes i tècniques que donen suport als processos de construcció i administració de tests. El seu objectiu últim, com veurem més endavant, és proporcionar les garanties científiques necessàries per a la mesura objectiva i estandarditzada dels fenòmens psicològics no observables a partir d'una mostra de comportaments. Aquest no és un objectiu menor i ha representat una contribució important de la psicologia a la teoria de la mesura desenvolupada en d'altres disciplines com la física. És a dir, d'acord amb les paraules atribuïdes a Galileo Galilei amb relació a la mesura científica: la manera en què la psicometria fa mesurables uns fenòmens psicològics que, per definició i en oposició als atributs físics, no són directament observables ni manipulables.

Començarem aquest viatge en el mòdul "Aproximació històrica i conceptes bàsics de la psicometria", en què situarem aquesta disciplina en el context general de la psicologia, revisant-ne els antecedents remots i més recents, presentant les aportacions més rellevants que han contribuït al seu establiment com a disciplina científica i oferint una definició formal que incorpori el paper dels tests en el marc de l'avaluació i la intervenció psicològiques. A continuació desenvoluparem els fonaments de la psicometria, partint d'una definició i classificació dels tests, progressant pels diferents models de mesura psicomètrica desenvolupats en les diferents teories dels tests, introduint breument la teoria clàssica dels tests i, finalment, recapitulant aquests fonaments amb la revisió del procés d'inferència psicomètrica. Conclourem el mòdul amb una discussió dels processos de construcció i administració de tests, presentant les

fases principals en què podem estructurar el disseny i la construcció d'un nou test, oferint alguns criteris importants per a l'avaluació de les característiques i la valoració de la conveniència dels tests disponibles en la literatura i, finalment, abordant els aspectes ètics i deontològics vinculats a l'ús de tests en la pràctica professional de la psicologia.

En el mòdul “Fiabilitat” tractarem de manera específica un aspecte fonamental per a la mesura dels fenòmens psicològics mitjançant tests: l'obtenció de puntuacions consistents i precises. Com passa amb qualsevol procés de mesura, el desenvolupament i l'administració de tests requereix el coneixement de l'error que es pot cometre. Si aquest error de mesura és gran, de manera que les puntuacions obtingudes no reflecteixen adequadament els fenòmens psicològics objecte d'interès, els tests no proporcionen la confiança necessària per a complir amb el seu objectiu principal al servei de l'avaluació psicològica. Així, en aquest segon mòdul abordarem la fiabilitat a partir de la perspectiva de la teoria clàssica dels tests, començant amb una descripció del model lineal clàssic, derivant el coeficient de fiabilitat i dedicant una atenció especial tant a la seva interpretació com a les diferents estratègies que s'han anat desenvolupant per a calcular-lo. A continuació es veuran tres factors importants que influeixen en la fiabilitat dels tests i es presentaran dos procediments per a estimar les puntuacions vertaderes dels subjectes a partir de les puntuacions obtingudes. Finalment, ens ocuparem de la fiabilitat dels tests referits a criteris, discutint-ne primer els trets principals i presentant a continuació els procediments clàssics disponibles per avaluar-ne la fiabilitat. Conclourem el mòdul amb una discussió dels mètodes més habituals en la determinació dels punts de tall que permeten la classificació correcta dels individus en referència al criteri.

El mòdul “Validesa” abordarà un altre aspecte clau per a la mesura indirecta dels fenòmens psicològics mitjançant tests: l'adequació als objectius per als quals han estat construïts i són utilitzats a la pràctica. Partint d'una revisió històrica, en aquest mòdul abordarem les diferents aproximacions que la psicometria ha anat proposant i definirem la validesa dels tests com el grau en què l'evidència empírica i la teoria donen suport a la interpretació de les puntuacions amb relació amb el seu ús específic. Com veurem a continuació, actualment la validesa no es considera com una propietat intrínseca dels tests sinó que és més aviat producte de l'anàlisi de la seva adequació al propòsit específic al qual serveixen. Per a fer-ho, els professionals interessats pel desenvolupament i l'administració de tests han de recollir i acumular les evidències científiques necessàries seguint diferents estratègies. Així, en aquest mòdul estructurarem els indicis de validesa dels tests en cinc grans grups: evidències basades en la validesa del contingut, basades en el procés de resposta, basades en l'estructura interna del test, basades en la relació amb d'altres mesures i, finalment, basades en les conseqüències de l'avaluació. Començarem definint cadascun d'aquests indicis de validesa i a continuació veurem en detall les diferents estratègies disponibles per a obtenir les evidències necessàries. Final-

ment, recapitularem aquesta discussió i veurem els factors que afecten la validesa dels tests tenint en compte la seva influència en aquestes estratègies per a obtenir els diferents indicis.

A continuació, en el mòdul “Transformació i interpretació de les puntuacions” ens ocuparem dels aspectes metodològics implicats en el tractament de la mesura indirecta dels fenòmens psicològics mitjançant tests. Més enllà d’algunes qüestions teòriques importants vinculades al procés de construcció de tests, en aquest mòdul discutirem les operacions que els professionals porten a terme per tal de fer interpretables les puntuacions obtingudes. Com veurem més endavant, aquestes puntuacions no són, *per se*, informatives i han de ser sempre interpretades per a fer-les útils d’acord amb el propòsit amb què els tests han estat desenvolupats. Així, començarem presentant el marc general d’interpretació de les mesures obtingudes mitjançant tests i abordarem algunes estratègies importants per a la transformació de puntuacions com són la construcció de percentils o de puntuacions estandarditzades i la utilització de normes cronològiques. Aquestes estratègies, que veurem en detall tenint en compte les seves característiques, la manera de calcular-les i les seves limitacions, serveixen per a recodificar les puntuacions obtingudes en un nou sistema de valors que en faciliti la interpretació sense afectar la diferent posició dels subjectes amb relació a les magnituds de les puntuacions originals. A continuació tractarem del procés de baremació o escalament de la mesura, que té per objectiu establir una connexió entre la puntuació de l’individu i l’execució observada en un grup de referència. Finalment, conclourem el mòdul amb una exposició de les diferents estratègies disponibles per a fer equiparables les puntuacions que proporcionen tests diferents que tenen per objectiu la mesura dels mateixos fenòmens psicològics.

En el mòdul “Anàlisi dels ítems” introduïrem breument un últim aspecte important per a la mesura indirecta dels fenòmens psicològics mitjançant tests: l’anàlisi del funcionament dels ítems que conformen els mateixos tests. Tot i que és una qüestió molt rellevant per a la psicometria, especialment en el disseny i la construcció de nous instruments, no sempre forma part dels programes de formació dels futurs psicòlegs. En aquest mòdul abordarem l’anàlisi de les propietats dels ítems en el cas específic de les proves d’execució màxima –també anomenades *tests d’habilitat o de potència*–, que tenen per objectiu avaluar la competència, l’aptitud o els coneixements dels individus a partir de l’encert o la qualitat de les seves respostes. Com veurem, a diferència de les proves d’execució típica, aquest tipus de proves discrimina respostes correctes i incorrectes, i és aquesta la base emprada per a puntuar les execucions individuals. Partint d’una definició d’aquests dos tipus de proves, començarem discutint algunes directives importants per a la construcció dels ítems, i presentarem una prova d’execució màxima fictícia que ens servirà per a il·lustrar aquesta exposició. Així, abordarem les propietats dels ítems d’acord amb la formulació de la teoria clàssica dels tests i discutirem els diferents procediments disponibles per a avaluar la dificultat i la discriminació dels ítems. Finalment, apuntarem la lògica proposada per la teoria de la resposta a l’ítem i conclou-

rem presentant el desenvolupament dels càlculs necessaris per a avaluar les propietats dels ítems que conformen la prova fictícia que hem utilitzat per a il·lustrar els diferents procediments.

Finalment, tancarem aquesta aproximació a la psicometria amb una selecció d'activitats pràctiques que permetran posar en joc els continguts desenvolupats en els diferents mòduls. Partint d'una exposició de tres casos ficticis basats en l'ús de tests, es plantejaran diferents exercicis vinculats a la fiabilitat, la validesa, la transformació i interpretació de les puntuacions i l'anàlisi dels ítems. Aquests exercicis, juntament amb el seu desenvolupament i les solucions corresponents, tenen per objectiu facilitar l'assoliment dels objectius plantejats al llarg del text. També es pot trobar un annex amb les taules de distribucions necessàries per a portar a terme aquestes activitats pràctiques.

Objectius

Aquesta assignatura té com a objectiu general:

1. Conèixer els fonaments de la psicometria com a disciplina científica encarregada de la mesura indirecta dels fenòmens psicològics mitjançant el desenvolupament i l'administració de tests.

A banda de l'objectiu general, té com a objectius específics:

1. Situar la psicometria en el context general de la psicologia al servei de l'avaluació i la intervenció psicològiques.
2. Saber definir i classificar els diferents tipus de tests disponibles.
3. Conèixer els diferents models de mesura desenvolupats per la psicometria a les diferents teories dels tests.
4. Entendre el procés d'inferència psicomètrica i conèixer els reptes específics que la mesura mitjançant tests ha d'abordar en el context del mètode científic.
5. Conèixer les implicacions pràctiques dels processos de construcció i administració de tests, fent un èmfasi especial en els aspectes ètics i deontològics vinculats amb el seu ús.
6. Entendre el concepte de *fiabilitat* des de la perspectiva psicomètrica.
7. Saber calcular i interpretar els coeficients de fiabilitat des de la perspectiva de la teoria clàssica dels tests.
8. Conèixer els factors que afecten la fiabilitat d'una mesura.
9. Saber estimar les puntuacions vertaderes dels subjectes a partir de les seves puntuacions empíriques.
10. Conèixer els procediments per a abordar la fiabilitat dels tests referits a criteri.
11. Conèixer els mètodes disponibles per a determinar el punt de tall per a classificar els individus.

- 12.** Conèixer els processos de validació dels tests que permeten inferir-ne l'adequació als objectius per als quals han estat construïts i són utilitzats a la pràctica.
- 13.** Saber definir i classificar els tipus de validesa en funció dels diferents indicis que es poden recollir com a evidències.
- 14.** Conèixer de manera pràctica les diferents formes de validesa per a saber si les conclusions que es treuen a partir de l'aplicació dels tests resulten adequades.
- 15.** Conèixer els factors que afecten els diferents tipus d'indis de validesa.
- 16.** Saber escollir el test més adequat en funció dels indicis disponibles de la seva validesa.
- 17.** Desenvolupar un punt de vista crític en la interpretació de les puntuacions obtingudes mitjançant tests.
- 18.** Conèixer les diverses estratègies disponibles per a transformar i interpretar les puntuacions dels tests.
- 19.** Conèixer les diverses estratègies disponibles per a equiparar les puntuacions obtingudes amb diferents instruments que mesuren els mateixos fenòmens psicològics.
- 20.** Entendre què és un barem i quins són els trets fonamentals que li proporcionen qualitat.
- 21.** Conèixer els diferents procediments disponibles per a valorar els ítems de les proves d'execució màxima.
- 22.** Conèixer les directives disponibles per a la construcció d'ítems que conformen les proves d'execució màxima.
- 23.** Saber valorar l'adequació dels ítems de les proves d'execució màxima a partir de les seves propietats psicomètriques.
- 24.** Conèixer les diferències bàsiques en l'anàlisi dels ítems des de les perspectives de la teoria clàssica dels tests i la teoria de resposta a l'ítem.

Continguts

Mòdul didàctic 1

Aproximació històrica i conceptes bàsics de la psicometria

Julio Meneses

1. La psicometria en el context de la psicologia
2. Fonaments de la psicometria
3. Construcció i administració de tests

Mòdul didàctic 2

Fiabilitat

Maite Barrios i Antoni Cosculluela

1. Concepte de fiabilitat segons la teoria clàssica
2. Equivalència de les mesures: mètode de les formes paral·leles
3. Estabilitat de les mesures: mètode test-retest
4. Consistència interna
5. Factors que afecten la fiabilitat
6. Estimació de la puntuació vertadera
7. Fiabilitat dels tests referits al criteri
8. Estimació dels punts de tall

Mòdul didàctic 3

Validesa

Luis Manuel Lozano i Jaume Turbany

1. Què és la validesa
2. Evidència de validesa basada en el contingut
3. Evidència de validesa basada en el procés de resposta
4. Evidència de validesa basada en l'estructura interna
5. Evidència de validesa basada en la relació amb altres variables
6. Evidència de validesa basada en les conseqüències de l'aplicació
7. Factors que afecten la validesa

Mòdul didàctic 4

Transformació i interpretació de les puntuacions

Sergi Valero

1. Interpretació d'una puntuació
2. Transformació de les puntuacions
3. Baremació
4. Equiparació de puntuacions

Mòdul didàctic 5

Anàlisi dels ítems

Albert Bonillo

1. Tipus de proves
2. Directives en la construcció d'ítems

3. Teoria clàssica
4. Teoria de resposta a l'ítem

Mòdul didàctic 6

Activitats pràctiques

Antoni Cosculluela

1. Introducció
2. Enunciats dels exercicis sobre fiabilitat
3. Enunciats dels exercicis sobre validesa
4. Enunciats dels exercicis sobre transformació i interpretació de les puntuacions
5. Enunciats dels exercicis sobre anàlisi dels ítems
6. Solucionari dels exercicis sobre fiabilitat
7. Solucionari dels exercicis sobre validesa
8. Solucionari dels exercicis sobre transformació i interpretació de les puntuacions
9. Solucionari dels exercicis sobre anàlisi dels ítems

Mòdul didàctic 7

Activitats pràctiques. Solucions utilitzant el programa Excel

Maite Barrios i Antoni Cosculluela

1. Introducció als aspectes bàsics del programa Excel
2. Solucions dels exercicis de fiabilitat
3. Solucions dels exercicis de validesa
4. Solucions dels exercicis de transformació i interpretació de puntuacions
5. Solucions dels exercicis d'anàlisi d'ítems

Annex. Taules de distribució

Julio Meneses

Bibliografia

Més enllà de les referències citades en els diferents mòduls, que serveixen per a aprofundir en alguns aspectes que van més enllà dels límits d'aquest text, a continuació oferim una selecció de contribucions desenvolupades en el nostre context immediat que poden ser d'utilitat per a obtenir una visió complementària a aquesta aproximació a la psicometria.

Fernández-Ballesteros, R. (1997). Evaluación psicológica y tests. A: A. Cordero (ed.). *La evaluación psicológica en el año 2000* (pp. 11-26). Madrid: TEA Ediciones.

El llibre editat per TEA entorn dels reptes per a l'avaluació psicològica ofereix algunes lectures interessants sobre el paper que tenen els tests en l'exercici professional dels psicòlegs. En aquest sentit, el capítol de la professora Fernández-Ballesteros presenta una síntesi excel·lent sobre els reptes de la mesura indirecta dels fenòmens psicològics mitjançant tests abordant una breu aproximació històrica, una discussió sobre els usos terminològics, l'encaix dels tests en el procés general de l'avaluació psicològica i una discussió sobre la validesa de les puntuacions dels tests.

Yela, M. (1996). Los tests. *Psicothema*, 8, suplem. 1, 249-263. Disponible en línia a <http://www.psicothema.com/pdf/660.pdf>.

Aquest és un text clàssic publicat l'any 1987 pel professor Yela en el seu manual *Introducción a la teoría de los tests*, i reproduït en els suplementes de la revista *Psicothema*, en què presenta una definició i descripció general dels tests com a reactius que revelen o donen testimoni fidel dels fenòmens psicològics no observables. El text també ofereix una aproximació històrica al seu desenvolupament, una classificació dels diferents tipus de tests disponibles i, finalment, una discussió sintètica de les diferents fases implicades en el desenvolupament de nous instruments.

Muñiz, J. (1998). La medición de lo psicológico. *Psicothema*, 10(1), 1-21. Disponible en línia a <http://www.psicothema.com/pdf/138.pdf>.

Aquest text correspon a la conferència inaugural del curs 1997-1998 de la Universitat d'Oviedo, on vam tenir l'oportunitat d'escoltar la veu autoritzada del professor Muñiz amb relació als importants reptes que planteja la mesura dels fenòmens psicològics no observables per a la psicologia. L'article comença amb una descripció de les característiques essencials d'aquests fenòmens, aborda els orígens de la mesura mitjançant tests i discuteix tres propietats bàsiques per a un ús adequat dels tests: la fiabilitat, la validesa i la fonamentació teòrica. Amb relació a aquesta última propietat, la seva exposició dels diferents models de mesura psicomètrica proposats per Fraser (1980) és una aproximació alternativa interessant a la que hem desenvolupat en el nostre text.

Muñiz, J. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. *Papeles del Psicólogo*, 31(1), 57-66. Disponible en línea a <http://www.papelesdelpsicologo.es/pdf/1796.pdf>.

Aquest número de la revista *Papeles del Psicólogo* és una referència important per als professionals interessats per la mesura indirecta dels fenòmens psicològics mitjançant tests. Entre els seus articles, la contribució de Muñiz ofereix una aproximació molt entenedora a les teories dels tests i el paper que tenen en l'establiment de les inferències a partir de les puntuacions obtingudes. Aquesta exposició arrenca amb una nota històrica que condueix cap a una caracterització de la teoria clàssica dels tests i una discussió de les seves limitacions, i finalment presenta les solucions que la teoria de resposta a l'ítem ha proposat recentment per fer front a aquestes limitacions. Més enllà de l'amença exposició, el lector interessat a conèixer aquestes dues aproximacions no hauria de deixar de tenir present la magnífica taula amb què, en la part final de l'article, es comparen les seves característiques bàsiques.

Prieto, G. i Delgado, A. R. (2010). Fiabilidad y validez. *Papeles del Psicólogo*, 31 (1), 67-74. Disponible en línea a <http://www.papelesdelpsicologo.es/pdf/1797.pdf>.

Al mateix número de la revista, els professors Prieto i Delgado ofereixen una panoràmica de la fiabilitat i la validesa dels tests, tant des d'un punt de vista conceptual com atenent als procediments més habituals per a la seva avaluació. L'article discuteix algunes nocions errònies sobre aquests principis, com són considerar la fiabilitat i la validesa com a propietats intrínseques dels tests o considerar-les de manera absoluta i no com una qüestió de grau.

Ferrando, J. i Anguiano-Carrasco, C. (2010). El análisis factorial como técnica de investigación psicológica. *Papeles del Psicólogo*, 31(1), 18-33. Disponible en línea a <http://www.papelesdelpsicologo.es/pdf/1793.pdf>.

Una tercera contribució interessant publicada en aquest número de la revista *Papeles del Psicólogo* és la dels professors Ferrando i Anguiano-Carrasco, on proposen una aproximació molt assequible a l'anàlisi factorial com a instrument d'investigació psicològica. Després d'una revisió conceptual, els autors discuteixen les diferències principals entre l'anàlisi factorial exploratori i l'anàlisi factorial confirmatori, i presenten els procediments principals implicats per a estimar els models corresponents. A continuació, il·lustren les diferents etapes implicades en una investigació, des del seu disseny i la recollida de dades fins a la interpretació de la solució final. Es presenta també el programa Factor, un recurs de distribució lliure molt interessant per a portar a terme tots els càlculs implicats en l'anàlisi factorial.

Moreno, R., Martínez, R. J., i Muñiz, J. (2004). Directrices para la construcción de ítems de elección múltiple. *Psicothema*, 16(3), 490-497. Disponible en línea a <http://www.psicothema.com/pdf/3023.pdf>.

Aquest és un article molt interessant per als professionals interessats pel desenvolupament dels ítems d'elecció múltiple que conformen un nou test orientat a l'avaluació de competències, aptituds o coneixements. A partir d'una revisió crítica de les diferents directrius existents, els professors Moreno, Martínez i Muñiz es proposen unificar la diversitat de criteris i proposen un conjunt de recomanacions que, sota el principi de parsimònia, faciliti una redacció adequada d'aquest tipus d'ítems. Es presenten un total de dotze directrius bàsiques que els autors, a més a més, il·lustren amb alguns exemples.

Prieto, G. i Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo*, 77, 65-75. Disponible en línia a <http://www.papelesdelpsicologo.es/vernumero.asp?id=1102>.

Per la seva banda, els professors Prieto i Muñiz presenten en aquest article el model d'avaluació dels tests desenvolupat pel Colegio Oficial de Psicólogos – actualment, Consejo General de Colegios Oficiales de Psicólogos –, on proposen un procediment d'avaluació de la qualitat i les característiques dels tests i presenten un qüestionari per a sistematitzar tot el procés. Aquest model ja ha estat utilitzat amb èxit per a avaluar deu dels tests més utilitzats pels psicòlegs espanyols i proporciona una guia interessant per a conèixer les decisions que els professionals interessats per la construcció de tests han de prendre. Aquest model també és una referència important per a la pràctica professional a l'hora de valorar la conveniència dels tests que hi ha en la literatura amb relació als objectius de l'avaluació psicològica.

Muñiz, J. i Fernández-Hermida, J. R. (2010). La opinión de los psicólogos españoles sobre el uso de los tests. *Papeles del Psicólogo*, 31(1), 108-121. Disponible en línia a <http://www.papelesdelpsicologo.es/pdf/1801.pdf>.

En aquest article els professors Muñiz i Fernández-Hermida presenten una part dels resultats d'un estudi conduït per l'European Federation of Psychologists' Associations que analitza les opinions sobre l'ús professional dels tests psicològics espanyols. Aquests resultats, basats en les anàlisis de les respostes de 3.126 professionals de la psicologia clínica, educativa i del treball, ens dona una fotografia interessant de la seva actitud general envers l'ús dels tests com a instruments d'avaluació psicològica a Espanya, i permet posar de manifest algunes limitacions importants que han de ser resoltes en el futur.

Lang, F. (2009). El principio de responsabilidad. *Papeles del Psicólogo*, 30(3), 220-234. Disponible en línia a <http://www.papelesdelpsicologo.es/pdf/1751.pdf>.

Aquest text forma part d'un altre número interessant de la revista *Papeles del Psicólogo* dedicat a la discussió de les qüestions ètiques i deontològiques vinculades a la pràctica professional dels psicòlegs. En l'article, el coordinador del Comitè d'Ètica de l'European Federation of Psychologists' Associations proposa una discussió de la responsabilitat professional envers les persones, les comunitats i la societat en general d'acord amb els principis del seu Meta-code of ethics. Aquest article és una versió ampliada del capítol amb què Lang va

contribuir al llibre *Ethics for European psychologists* i dóna l'oportunitat de reflexionar entorn d'uns exemples concrets que il·lustren els diferents aspectes en què es concreta el principi general de responsabilitat. Amb relació al que ens ocupa en aquesta aproximació a la psicometria, són d'especial rellevància els exemples 4, 7 i 8, en què planteja algunes consideracions importants amb relació a les conseqüències derivades de l'ús professional dels tests com a instruments d'avaluació psicològica.

International Test Commission (2000). International guidelines for test use. Disponible en línia a <http://www.intestcom.org/upload/sitefiles/41.pdf>.

En l'àmbit dels codis deontològics professionals dels psicòlegs, aquestes directrius representen una important contribució desenvolupada per la International Test Commission amb l'objectiu de donar una estructura coherent sota la qual es puguin entendre i aplicar els diferents codis i estàndards nacionals que desenvolupen els aspectes ètics i deontològics que afecten l'ús de tests. Més enllà de l'interès del seu articulat per a una pràctica professional responsable, aquestes directrius posen de manifest la importància del desenvolupament i l'adquisició de les competències necessàries per a portar a terme l'administració de tests, la interpretació i comunicació adequades dels resultats, i la resolució de les dificultats, malentesos i conflictes que se'n puguin derivar. Disposen d'una versió traduïda al castellà que aquesta organització proporciona gratuïtament per encàrrec per mitjà de la seva pàgina web.

Aproximació històrica i conceptes bàsics de la psicometria

PID_00216191

Julio Meneses

Temps mínim de dedicació recomanat: 4 hores



**Julio Meneses**

Professor dels Estudis de Psicologia i Ciències de l'Educació de la Universitat Oberta de Catalunya (UOC) i investigador de l'Internet Interdisciplinary Institute (IN3). Llicenciat en Psicologia per la Universitat d'Oviedo, es va doctorar en el programa de Societat de la Informació i el Coneixement de la UOC i, actualment, treballa com a responsable de la metodologia en diversos projectes d'investigació a les ciències socials. La seva recerca s'orienta a l'aplicació dels mètodes de recerca quantitativa i a l'ús de tècniques d'anàlisi multivariant, particularment en les àrees de la introducció de les TIC als sistemes educatius, la relació dels infants i els joves amb la tecnologia i l'abandonament i el reingrés dels estudiants universitaris.

Índex

Introducció.....	5
1. La psicometria en el context de la psicologia.....	7
1.1. Una aproximació històrica a la psicometria	7
1.2. La psicometria avui	12
1.3. La psicometria en el context de l'avaluació psicològica	15
2. Fonaments de la psicometria.....	18
2.1. Definició i classificació dels tests	18
2.2. Models de mesura psicomètrica	22
2.3. Teoria clàssica dels tests	26
2.4. El procés d'inferència psicomètrica	29
3. Construcció i administració de tests.....	33
3.1. El procés de construcció de tests	33
3.2. Criteris per a la valoració de tests	36
3.3. Aspectes ètics i deontològics en l'ús de tests	39
Bibliografia.....	47

Introducció

Aquesta introducció a la psicometria es proposa donar a l'estudiant algunes claus importants per a abordar la complexitat dels coneixements, procediments i valors vinculats amb el desenvolupament i l'administració de tests. Partint de l'enquadrament de la psicometria en el context general de la psicologia, començarem desenvolupant una aproximació històrica al seu naixement, des dels antecedents remots fins als primers desenvolupaments, descriurem l'estatus actual de la psicometria com a disciplina científica i, finalment, construirem una definició formal tenint en compte el paper que tenen els tests com a instruments d'avaluació psicològica. A continuació abordarem els fonaments de la psicometria, en què tindrem l'oportunitat d'oferir una definició i classificació dels tests, abordar els models de mesura psicomètrica, introduir la teoria clàssica dels tests i recapitular el procés d'inferència psicomètrica en què es basen els tests. Finalment, discutirem les qüestions relatives a la construcció i administració de tests, tractant les diferents fases implicades en el desenvolupament i oferint alguns criteris importants per a la valoració dels tests disponibles en la literatura. Conclourem aquesta introducció amb una discussió dels aspectes ètics i deontològics vinculats amb la utilització de tests en la pràctica professional de la psicologia.

1. La psicometria en el context de la psicologia

Entendre un camp d'estudi determinat implica, en la majoria dels casos, conèixer-ne les arrels fundacionals. És a dir, conèixer els problemes a què es van enfrontar els pioners, la manera en què els entenien i les solucions que hi van donar. Tot i que aquest no és el lloc per a fer un viatge en profunditat¹, una aproximació històrica a la psicometria ha de començar recollint el que es consideren els seus orígens i antecedents. Malgrat que el desenvolupament i l'administració de tests psicològics és una pràctica desenvolupada fonamentalment a partir del segle XX, és possible trobar alguns antecedents remots en cultures tan antigues com la xinesa. Partint d'aquests antecedents, i repassant algunes de les contribucions més importants al seu desenvolupament com a disciplina científica, estarem en disposició d'oferir una definició formal de la psicometria que ens permeti situar-la en el context general de la psicologia.

⁽¹⁾Entre d'altres, vegeu els treballs clàssics de Goodenough (1949), DuBois (1970), i Hilgard (1987), més recentment, Buchanan i Finch (2005), o Jones i Thissen (2007).

1.1. Una aproximació històrica a la psicometria

Tal com s'acostuma a assenyalar, el desenvolupament de les primeres dinasties de l'**antic imperi xinès** va generar els primers sistemes d'avaluació dels individus en funció de la seva habilitat. Tot i que alguns cronistes han apuntat a referències tan antigues com les de l'any 2000 aC², l'estudi de les evidències arqueològiques ha posat en dubte aquesta antiguitat. En qualsevol cas, tal com ha apuntat Bowman (1989), les proves documentals que es conserven permeten situar aquest origen en un període relativament més recent, durant la dinastia Tang (618-907). Durant aquests anys es va desenvolupar un sistema d'avaluació imperial que va permetre la selecció i promoció dels funcionaris dels diferents departaments de l'Administració. Aquest sistema va tenir un important impuls durant la dinastia Ming (1368-1644) quan es va establir un examen institucional segons el mèrit per a tots els funcionaris dels diferents nivells territorials –des del nivell municipal al nacional–, posant en marxa un dels primers sistemes de classificació oficial mitjançant l'expedició dels primers títols formals que acreditaven l'accés als diferents nivells de responsabilitat.

⁽²⁾Per exemple, podeu veure DuBois (1970).

Aquests sistemes d'avaluació són fàcilment homologables als que l'**educació formal europea**, especialment la universitària, va desenvolupar a partir de la introducció dels exàmens orals per als seus estudiants a partir del segle XIII. Tal com assenyala Rogers (1995), la invenció i la incorporació del paper en la vida quotidiana a partir del segle XVI va facilitar el trànsit cap a les proves escrites que durant el segle XIX van esdevenir els primers mecanismes de selecció competitiva dels estudiants universitaris. Com a antecedents remots de la psicometria, tant l'avaluació en l'educació formal com l'establiment del sistema d'avaluació imperial van portar un canvi important en la concepció sobre el

judici de les capacitats de les persones. D'aquesta manera, es va anar traslladant la confiança en el judici personal basat en impressions cap a l'administració de proves institucionals bassades en una autoritat imparcial que objectivava les destreses i coneixements requerits en els àmbits educatiu i administratiu.

Tot i que algunes vegades s'ha passat per alt, un altre antecedent remot es trobaria a l'inici de l'**avaluació psiquiàtrica** a mitjan segle XIX. D'acord amb el treball de Bondy (1974), és convenient tenir present l'esforç dels primers professionals orientats a l'estudi dels problemes mentals i les lesions cerebrals en l'establiment del que es podrien considerar les primeres proves d'avaluació psicològica. Així, per exemple, es van desenvolupar els primers tests per a avaluar les conseqüències del dany cerebral, algun d'ells tan elaborat que exigia dedicar-hi períodes de 100 hores. Malgrat les seves limitacions, entre d'altres per l'absència de procediments estàndards en el seu ús, aquestes primeres proves van incorporar molts dels elements que, amb l'evolució de l'estudi dels símptomes psicològics, encara són utilitzats en les proves de diagnosi actuals.

Desenvolupades les bases per a l'examen individual, tant en contextos d'activitat quotidiana com en situacions de trastorn psicològic o accident, els antecedents recents de la psicometria es trobarien en el desenvolupament de l'**estudi sistemàtic de les diferències humanes** durant el segle XIX. Primer gràcies als treballs de Friedrich W. Bessel (1784-1846) i Carl F. Gauss (1777-1855) –que van ser pioners en l'estudi de les diferències individuals en la percepció en el camp de l'astronomia– i després a les contribucions de Gustav T. Fechner (1801-1887) i Hermann von Helmholtz (1821-1894) en el desenvolupament de la psicofísica –que va representar l'inici de la psicologia com a disciplina acadèmica–, totes dues aproximacions van constituir un avenç important en la sistematització de la mesura de les sensacions psicològiques produïdes per l'estimulació física (Boring, 1978, per a una revisió històrica). Així, mitjançant el desenvolupament de les primeres lleis que relacionen estímuls físics i sensacions, es van anar assentant les bases per a la mesura psicològica. D'aquesta manera, es van perfeccionar els mètodes de presentació dels estímuls i el registre de les respostes, es va treballar en la millora de la precisió de les mesures, i es van adoptar condicions controlades per a la seva consecució. Tots ells elements indispensables per al plantejament posterior dels problemes més importants que la psicometria hauria d'afrontar en el seu desenvolupament inicial.

Així, una vegada apuntats els antecedents remots i més recents, hem d'esmentar els treballs de Sir Francis Galton (1822-1911), James McKeen Cattell (1860-1944) i Alfred Binet (1857-1911) com els veritables **pioners de la psicometria moderna**. Com a continuadors de les aproximacions anteriors a l'estudi sistemàtic de les diferències humanes, Galton i Cattell van contribuir necessàriament a l'establiment de la psicologia experimental a Europa i els Estats Units, respectivament, mitjançant la creació dels primers laboratoris antropomètrics per a l'estudi de les característiques humanes (Valentine, 1999, i Sokal, 1982) i el desenvolupament dels primers tests per a l'avaluació de les diferències sensorials, perceptives i de comportament. Obrint un nou camí per

a l'estudi científic de la psicologia, tots dos autors es van aventurar a postular relacions entre les seves mesures i l'intel·lecte, i van arribar a conclusions a vegades controvertides i d'altres àmpliament criticades amb posterioritat pel seu simplisme. Per la seva banda, Binet va adoptar un enfocament innovador i va ser responsable del que es considera el primer test d'aplicació general per a la mesura de les habilitats cognitives.

Responent a la necessitat d'identificar els estudiants amb problemes per a assolir els objectius de la instrucció ordinària a les escoles de París del principi del segle XX, Binet i el seu col·lega Théodore Simon van recollir els incipients moviments en l'avaluació de la discapacitat cognitiva mitjançant bateries de tests (Nicolas i Ferrand, 2002). A partir d'aquests, i a petició del Ministeri d'Educació francès, el 1905 van desenvolupar la primera prova que va permetre classificar els infants segons la intel·ligència (Wolf, 1969). Aquest treball va permetre el naixement de l'interès de la psicologia per l'atenció a les necessitats educatives especials, però no va ser fins al 1908 quan aquests autors van publicar una revisió de la seva escala que va permetre mesurar el que es va anomenar *el nivell mental*. Gràcies a l'adopció d'un grup de referència compost per nens que tenien entre tres i tretze anys, i una vegada ordenats els ítems en funció de l'edat en què eren típicament resolts, van desenvolupar el primer test per a quantificar la intel·ligència. Aquesta quantificació, prenent com a referència l'edat biològica de l'infant, va conduir a la definició del que es coneix, fins a l'actualitat, com *l'edat mental*.

Adoptant el mateix enfocament orientat a la pràctica, la incipient psicometria es va veure impulsada en els circuits acadèmics entorn del debat sobre la **mesura de la intel·ligència** mitjançant el desenvolupament de tests (Martin, 1997). A partir dels treballs de Binet en el camp de l'educació, l'interès per l'avaluació de la cognició va fer el salt al circuit internacional de la mà de Lewis M. Terman (1877-1956), el qual va publicar el 1916 la revisió Stanford-Binet. D'acord amb el relat de Goodenough (1949), aquesta contribució va establir una fita important en el desenvolupament dels tests tal com els coneixem actualment. Revisant els ítems originals, incorporant pautes clares d'administració i destacant la importància de la representativitat de les mostres per a la interpretació correcta de les puntuacions, va esdevenir l'escala de referència per a la mesura de la intel·ligència durant les dècades següents. De fet, va fer un altre pas important per al desenvolupament de la psicometria, en la mesura en què es va estendre l'ús d'aquesta escala a l'exèrcit nord-americà durant la Primera Guerra Mundial. En aquest context, i gràcies al suport del Govern dels Estats Units, la psicometria va créixer en el centre de les polítiques militars³.

⁽³⁾Entre d'altres, podeu veure Zeidner i Drucker (1988).

Recuperant l'esperit classificador, els tests haurien de servir com a instruments d'avaluació psicològica per al reclutament dels nous soldats. Així, a partir de la generalització de l'ús de la revisió Stanford-Binet, es van desenvolupar nous tests i es va desenvolupar l'administració en grup per fer més eficient el procediment de mesura. D'aquests esforços, entre d'altres, cal destacar l'empremta de Robert Yerkes (1876-1956) en el desenvolupament dels tests Army Alpha i

Beta (Carson, 1993), que també van abordar una controvèrsia important entorn dels **biaixos culturals dels tests** que es va anar construint durant aquestes primeres dècades del segle XX. Tal com es va discutir àmpliament (Jensen, 1980), les noves proves per a mesurar la intel·ligència podien no estar lliures d'influències culturals de manera que subestimarien els que no parlessin la llengua anglesa, els analfabets i qui patís alguna discapacitat visual o auditiva. Recollint els incipients moviments entorn de les proves no verbals, la versió beta del test d'intel·ligència de l'exèrcit nord-americà va representar el reconeixement de la importància d'aquestes diferències, com també la necessitat de minimitzar-les en qualsevol context en què s'utilitzin els tests com a instruments d'avaluació psicològica.

Una qüestió interessant, per les conseqüències que va tenir més endavant en el desenvolupament de la psicometria, va ser la concepció que es va anar construint entorn del mateix concepte d'*intel·ligència*. Arrelada a les nocions biològiques i hereditàries dels pioners, i sustentada en els desenvolupaments estadístics de Charles Spearman (1863-1945), un corrent important va concebre la intel·ligència com un únic factor –el factor *g*, segons Spearman– que explicaria les puntuacions en els diferents tests d'intel·ligència desenvolupats fins aleshores. En canvi, des de la posició “alternativa” que va encapçalar Louis Leon Thurstone (1887-1955), la intel·ligència estaria composta en realitat per diversos factors específics. Els seus avenços en el desenvolupament estadístic de les **tècniques d'anàlisi factorial** que va introduir Spearman van permetre comprendre millor aquesta aparent contradicció⁴. De fet, es podia entendre el factor general d'intel·ligència com una mena de factor subjacent que recolliria la variabilitat de les aptituds específiques de Thurstone, tant si eren relatives a la comprensió verbal com a la competència numèrica o a l'habilitat espacial, entre d'altres. De la mateixa manera, una vegada controlat l'efecte del factor general, el plantejament d'uns factors grupals per part de Spearman permetria recollir la variabilitat comuna observada en tests que comparteixen demandes verbals, numèriques o espacials similars. Més enllà de les conseqüències teòriques sobre la mesura de la intel·ligència, el perfeccionament de l'anàlisi factorial va permetre pavimentar el camí estadístic per al desenvolupament de la psicometria fins al final del segle XX.

⁽⁴⁾Per exemple, podeu veure Cattell (1943).

A més a més, aquesta visió de la intel·ligència com un conjunt determinat d'aptituds cognitives específiques va contribuir necessàriament al desenvolupament dels **primers estudis sistemàtics de validesa**. Tal com va passar durant la Primera Guerra Mundial, el Govern nord-americà va tornar a incorporar aquestes pràctiques als seus programes de reclutament uns quants anys abans de participar en la Segona Guerra Mundial. Tot i que els esforços anteriors havien donat els seus fruits, l'exèrcit nord-americà va trobar en aquest plantejament la manera de resoldre algunes limitacions de l'enfocament unitari de la intel·ligència, principalment per la seva escassa capacitat per a seleccionar candidats amb perfils molt especialitzats. Va ser durant aquest període quan, sota la direcció de John C. Flanagan (1906-1996), les forces d'aviació van administrar un conjunt de bateries de tests per a seleccionar i classificar

els pilots, enginyers de vol i altres tècnics encarregats dels instruments de navegació. L'estudi de la relació entre les puntuacions en les diferents aptituds i l'èxit en la formació posterior dels reclutes va ser clau en aquest procés (Goslin, 1963), permetent la millora dels procediments de selecció i classificació, i conformant un dels primers estudis empírics de la validesa de les mesures psicològiques.

No hem d'oblidar, però, el treball que paral·lelament es va a portat a terme amb relació a l'**avaluació de la personalitat**. Encara en el context de les polítiques militars de la Primera Guerra Mundial, el Govern dels Estats Units es va enfrontar a un altre tipus de problema pràctic. Més enllà de la selecció i classificació dels reclutes, un important esforç es va adreçar a la identificació de candidats susceptibles de patir trastorns psicològics. Robert S. Woodworth (1869-1962) va ser l'encarregat de desenvolupar un nou tipus de prova que permetés avaluar l'estabilitat emocional dels soldats per a minimitzar-ne la presència a les seves files. Aquest test, el seu *Personal Data Sheet*, va introduir un conjunt de preguntes amb resposta positiva o negativa que, a diferència de l'avaluació de la intel·ligència, no contenia respostes necessàriament correctes o incorrectes. D'aquesta manera, Woodworth va elaborar una prova que permetia detectar els casos problemàtics, no ja amb la comparació de la resposta individual amb una mostra representativa de la població, sinó a partir de les respostes donades per individus amb trastorns psicològics ja diagnosticats. Aquest treball va assentar les bases per al desenvolupament dels tests d'avaluació de la personalitat posteriors (Gibby i Zickar, 2008), com també va començar a definir les bases per a tractar adequadament la possibilitat de frau en la resposta, tant per a amagar com per a fer veure que es pateix un trastorn.

El prolífic Thurstone, que va desenvolupar les primeres proves d'anàlisi de consistència interna dels tests en l'àmbit de l'avaluació de la personalitat, va fer una altra contribució important al desenvolupament de la psicometria. Partint dels experiments desenvolupats pels pioners (Gulliksen, 1968), especialment en la branca de la psicofísica, el 1927 va proposar la llei del judici comparatiu com el mètode per a la **mesura de les actituds, les preferències i els valors**. Aquesta innovació va consistir a traslladar el judici perceptiu sobre parelles d'estímul físics –per exemple, la tria entre dos sons segons la intensitat– a la valoració de característiques psicològiques no estrictament relacionades amb les propietats físiques –per exemple, la tria entre dos comportaments segons l'acceptabilitat. Després de proposar una segona llei, la llei del judici categòric, l'aproximació de Thurstone a la mesura de les actituds estava preparada per a ser aplicada al judici general, és a dir, sense requerir la comparació entre parelles de fenòmens. D'aquesta manera, tal com pretenien Thurstone i els seus continuadors (Bock i Jones, 1968), la psicometria disposaria d'un mètode d'escalament per a mesurar i tractar numèricament valoracions individuals subjectives en la cerca d'una representació objectiva dels fenòmens psicològics. Aquest mètode va donar un suport metodològic important a les dècades següents de recerca en psicologia social.

1.2. La psicometria avui

Tot i la dificultat de fer una aproximació històrica a partir de la selecció d'algunes de les aportacions més importants, aquest exercici ens permet il·lustrar el naixement d'una disciplina a partir dels problemes i les diferents aproximacions dels pioners. La psicometria, un nou espai de treball metodològic entorn del desenvolupament i l'administració de tests, es va anar conformant amb un component aplicat, orientada a les demandes en diferents contextos, i va formar part d'alguns dels debats teòrics més importants de la mateixa psicologia. No és, però, fins a la dècada dels anys trenta del segle XX que podem situar l'**inici de la seva constitució com a disciplina científica** tal com la coneixem actualment. És important començar assenyalant de nou el paper decisiu de Thurstone, el qual, amb l'objectiu d'establir i promoure la psicologia com una ciència quantitativa (Samejima, 2000), el 1935 va fundar la Psychometric Society i en va ser el primer president. A més a més, va ser l'impulsor de la primera revista especialitzada que encara és de referència obligada, la revista *Psychometrika*, que va anar publicant alguns dels treballs més importants sobre els quals es va formalitzar la psicometria. Pocs anys després, el 1946, va ser també el primer president de la Divisió d'Avaluació i Mesura de l'American Psychological Association.

En aquest sentit, podríem definir el període entre les dècades dels anys trenta i els seixanta com l'**època daurada de la psicometria**. És en aquest període que es publiquen, a més a més, els llibres i manuals més importants que la vertebrarien. Entre els manuals cal destacar *The reliability and validity of tests*, de Thurstone (1931), que sistematitzava el que s'havia desenvolupat fins aleshores amb relació a la teoria dels tests i suggeria el rol central de la fiabilitat com a requisit per a la validesa de les mesures en la psicometria. Tornarem més endavant sobre aquestes qüestions, elements bàsics de la mesura indirecta dels fenòmens psicològics mitjançant tests. En la mateixa època, també va ser important la primera edició del manual *Psychometric methods*, de Guilford (1936), un intent per organitzar el camp popi de la psicometria entorn de la teoria dels tests, l'escalament psicològic i el psicofísic. La teoria clàssica dels tests, com va ser àmpliament coneguda a partir dels treballs de Spearman sobre l'estimació dels errors de mesura, començava el seu camí en els circuits docents universitaris encarregats de formar els futurs psicòlegs.

D'altra banda, durant els anys trenta i especialment els quaranta, un dels debats més apassionants oberts per la psicometria va tenir lloc entorn del mateix concepte de *mesura*. Inspirat per la British Association for the Advancement of Science, psicòlegs i físics van proposar-se el repte de decidir si, amb els avenços en el desenvolupament dels tests, la mesura psicològica podia ser enquadrada dintre del model general de mesura dels atributs físics. En síntesi, mesurar consistia a quantificar, és a dir, determinar la quantitat en què un atribut és present en un objecte. I per a fer-ho, la mesura dels atributs físics depenia de la capacitat d'observació de relacions entre objectes com a conseqüència d'una operació o manipulació empírica. En tant que aquesta possibilitat només po-

Per saber-ne més

L'estudiant interessat a aprofundir sobre la constitució de la psicometria com a disciplina científica des d'una perspectiva històrica pot consultar les revisions de Hambleton (1994), Brennan (1997), Traub (1997) o Bock (1997).

dia ser satisfeta amb els objectes físics, la controvèrsia va tenir una resposta inicial contundent. Tal com van concloure Ferguson i els seus col·laboradors (1940), la mesura psicològica que es proposava la psicometria simplement no era possible, ja que aquest model no es podia estendre per a incloure els atributs psicològics que, per definició, no eren observables ni manipulables empíricament. El debat, però, no va quedar tancat i va ser Stevens (1946) amb el seu treball *On the theory of scales of measurement* qui va fer un pas fonamental per al desenvolupament del que posteriorment va ser anomenat el *model representacional de mesura psicomètrica*.

Stevens va definir la mesura com el procés d'assignació de números a objectes o esdeveniments d'acord amb unes regles, producte de les quals s'obtingrien els diferents tipus d'escalas proposades: nominal, ordinal, d'interval i de raó. Mesurar, en aquest sentit, no consistiria únicament en quantificar sinó que seria el producte de la utilització de diferents regles que, en darrer terme, determinarien el tipus d'operacions –o tècniques– estadístiques permeses en cada escala. No exempta de crítiques, és important destacar la importància de la seva contribució com un primer intent per superar les restriccions imposades per la quantificació dels atributs físics amb l'objectiu de resoldre la controvèrsia entorn de la mesura dels fenòmens psicològics. Així, el debat sobre els models o paradigmes de mesura va formar part important de l'agenda psicomètrica i es van posar les bases per al desenvolupament posterior d'altres models com l'operacional i el clàssic, sobre els quals tornarem més endavant. D'altra banda, tancat aquest breu repàs de les contribucions importants durant els anys quaranta, no podem deixar de fer menció de la publicació de Thurstone (1947) de l'influent *Multiple factor analysis*, que a partir dels treballs de Spearman, Kelley i Burt dona el suport estadístic necessari per a la construcció i validació dels tests durant les dècades següents.

En la dècada dels anys cinquanta assistim a la publicació de les altres dues obres de referència per a la psicometria moderna. Instal·lada en el circuit acadèmic, la feina portada a terme per diversos autors convergeix en el que s'ha anomenat la *teoria clàssica dels tests* (TCT). I ho fa de la mà de Gulliksen (1950) en el seu *Theory of mental tests*, en què formalitza el model lineal clàssic per primera vegada i en defineix les assumpcions principals. Com a teoria dels tests, la TCT proposa un nou enfocament basat en el concepte de puntuació vertadera. Partint de la puntuació empírica obtinguda mitjançant els tests, i seguint un conjunt determinat de supòsits, l'objectiu és descompondre-la en dues parts fonamentals per tal d'estimar l'error associat al procés de mesura i, així, inferir el valor real que es vol mesurar. No sense dificultats, es va convertir en poc temps en la teoria dels tests de referència i va estimular tant el debat sobre la mateixa mesura psicològica com el procés de desenvolupament dels tests durant les dècades següents. De manera anàloga, Torgerson (1958) publica *Theory and methods of scaling* i estableix el cànon per a l'escalament psicofísic i psicològic, és a dir, l'ordenació dels estímuls de manera paral·lela a la de les persones.

Finalment, els anys seixanta són una dècada de **revisió crítica i obertura de nous camins** per a la psicometria. A partir de la TCT, el debat entorn de l'estimació dels errors va cristal·litzar en dos nous corrents en el desenvolupament de tests: la teoria de la generalitzabilitat i la teoria de resposta a l'ítem. Per tal d'incrementar la precisió de les mesures, Cronbach i els seus col·laboradors⁵ van proposar la teoria de la generalitzabilitat, que mitjançant l'aplicació de l'anàlisi de variància permetria descompondre l'error genèric proposat per la TCT en la cerca dels seus diferents elements. Així, d'acord amb aquesta teoria, l'anàlisi de la fiabilitat es basa en el disseny d'investigacions que permeten analitzar les diferents fonts d'error –facetes, segons els seus termes– que afectarien el procés de mesura. Entre aquestes, per exemple, la forma del test, els ítems que el componen, les ocasions en què s'administra o els participants. Tot i que va ser una innovació important, la complexitat d'aquesta teoria va limitar la seva difusió a la pràctica i en poc temps els seus avenços van ser bàsicament reformulats per un segon enfocament, la teoria de resposta a l'ítem (TRI).

⁽⁵⁾Cronbach, Rajaratnam i Gleser (1963); Gleser, Cronbach i Rajaratnam (1965).

Així, la TRI es presenta com la resposta a les crítiques principals que havia rebut la TCT. En aquest sentit, és convenient fer esment de la publicació per part de Lord i Novick (1968) del seu *Statistical theories of mental test scores*. En aquest treball, també amb esperit de síntesi i anàlisi crítica de la feina desenvolupada en el context de la TCT, es plantegen les dificultats que aquest model no estava essent capaç de resoldre. Bàsicament, els problemes principals podrien ser organitzats entorn de la dependència dels instruments i les mostres utilitzades en el procés de construcció i administració dels tests. Durant les dècades posteriors la TRI ocuparà, juntament amb la TCT, un lloc privilegiat entre les teories dels tests desenvolupades per la psicometria. Tornarem més endavant sobre aquestes qüestions però val la pena tancar aquesta etapa daurada de la psicometria moderna assenyalant un últim desenvolupament que també va tenir el seu inici en la dècada dels anys seixanta. Es tracta dels tests referits a criteri que, en el context de l'educació, tenen com a objectiu avaluar la destresa de les persones en un camp de coneixement molt ben delimitat.

Avaluar en funció de criteris i no sobre la base de la norma –és a dir, escalar o ordenar els individus comparant-ne les puntuacions– no és un enfocament nou. Al contrari, com hem vist en el recorregut històric del naixement i establiment de la psicometria com a disciplina científica, aquesta forma d'avaluació ha estat objecte d'interès amb anterioritat, però és gràcies a les contribucions de Glaser (1963) i Popham i Husek (1969) que el treball en aquest context queda formalitzat. El que és innovador, des del punt de vista del desenvolupament dels tests, és l'enfocament en el procediment per a establir els estàndards d'avaluació i també en la consistència i precisió amb què es classifiquen els individus d'acord amb aquests estàndards. Tot i que va néixer en el marc de la TCT, aplicacions posteriors de la TRI han permès aprofitar els avenços en el desenvolupament dels tests referits a criteri com a instruments de mesura dels fenòmens psicològics.

1.3. La psicometria en el context de l'avaluació psicològica

Una vegada recorregut el camí dels antecedents i els orígens de la psicometria moderna, estem en disposició d'abordar-ne formalment la definició. I podem fer-ho a partir de les definicions que s'han anat proposant en els textos de referència. Aquesta tasca, però, no està exempta de risc. Els autors elaboren definicions concretes que, probablement, no reflecteixen tots els matisos amb què després les contextualitzen en els seus textos. Malgrat això, podem aprofitar aquestes propostes per a valorar els diferents aspectes en què es posa un èmfasi especial. Al cap i a la fi, no hi cap tot en una definició.

Més enllà del tòpic que diu que hi ha tantes definicions com autors, podem distingir tres grans aproximacions. Per una banda, un primer grup de definicions aborda la psicometria a partir dels **instruments que utilitza**. Des d'aquest corrent es circumscriu la psicometria com la disciplina encarregada de desenvolupar els fonaments per a la construcció i administració de tests (per exemple, vegeu Martínez Arias, Hernández Lloreda i Hernández Lloreda, 2006). Aquesta definició, tot i que curta i concisa, presenta algunes dificultats. En primer lloc, la paraula *test* és polisèmica, de manera que segons el context en què sigui emprada pot esdevenir sinònim d'altres termes amb significats diferents com *prova*, *examen* o fins i tot *assaig*. En segon lloc, més enllà de la precisió semàntica que es pot sobreentendre en el context de la psicologia, no deixa de ser en certa manera restrictiva. Si és ben cert que els tests psicològics són els instruments específics que desenvolupen i administren els psicòmetres, també ho és que són el resultat aplicat d'un procés més ampli caracteritzat pel desenvolupament de teories i mètodes orientats a la mesura indirecta dels fenòmens psicològics.

Un segon corrent defineix la psicometria a partir de l'**objecte a què sotmet o aplica el seu interès**. Així, la psicometria es defineix de manera més o menys genèrica com la disciplina científica orientada a l'avaluació o mesura dels fenòmens psicològics (per exemple, vegeu Rust i Golombok, 2009). Aquesta definició s'ajusta de fet al significat etimològic de la paraula *psicometria*, que en els seus orígens grecs podem trobar en la juxtaposició de les paraules *psique* – que significa 'ànima', 'alè' o 'intel·lecte' – i *metron* –en referència al procés de mesura. En aquesta línia es pot situar la definició de Kline (1998), que aborda específicament la psicometria com la tasca de desenvolupar mesures científiques fonamentals en les àrees de la personalitat i les capacitats. En aquest context, la científicitat es converteix en sinònim d'*estandardització* i, tot i no fer-hi referència explícita, al·ludeix indirectament a la utilització de tests en el context del mètode científic. Un altre aspecte implícit important és, amb relació a l'estatus mètric de les puntuacions obtingudes com a resultat del procés de mesura psicològica, el caràcter quantitatiu que hauria de tenir la psicometria com a disciplina científica. Com ja hem avançat en l'aproximació històrica, aquest és un tema controvertit sobre el qual haurem de tornar més endavant.

En una posició intermèdia, un tercer grup d'autors situa la psicometria en la **intersecció dels dos corrents anteriors**. En aquesta línia podem assenyalar per exemple la definició de Buchanan i Finch (2005), que especifica els dominis de mesura estandarditzada mitjançant tests amb relació a les habilitats, els atributs i les característiques psicològiques. És important assenyalar que, malgrat que en alguns casos es fa referència explícita a la utilització de tests, entre aquestes definicions podem trobar també un reconeixement explícit del sentit més ample del procés de mesura. Així, Holden (2000) defineix la psicometria com la teoria i la tècnica de mesura que, en el context de la psicologia, s'encarrega dels factors que són mesurables. En aquest sentit, Muñiz (2003) ho fa definint la psicometria com el conjunt de mètodes, tècniques i teories implicades en la mesura de les variables psicològiques tenint-ne en compte l'especialització en les propietats mètriques exigibles a aquest tipus de mesura. Finalment, Jones i Thissen (2007) descriuen la psicometria com la disciplina –quantitativa, especifiquen en aquest cas– encarregada de desenvolupar models i mètodes orientats principalment al resum, la descripció i l'establiment d'inferències a partir de les dades recollides en la investigació psicològica.

Com hem pogut observar, la definició de la psicometria s'ha desenvolupat fonamentalment entorn dels mitjans o instruments que utilitza –els tests– i l'objecte que persegueix amb el seu ús –la mesura dels fenòmens psicològics. En canvi, una aproximació comprensiva a aquesta definició hauria de considerar un tercer element important vinculat amb la **finalitat última a la qual serveix**. Tal com hem esmentat en aquest apartat, la psicometria adquireix tot el seu significat amb relació a l'àrea de la psicologia en què té el seu paper fonamental: l'avaluació psicològica. D'aquesta manera, hem de tenir present que la mesura dels fenòmens psicològics mitjançant el desenvolupament i l'administració de tests no és més que una part del procés general d'avaluació, i no sempre és la més important⁶.

⁽⁶⁾Podeu veure Fernández-Ballesteros (1997) per a una discussió més àmplia.

La psicometria, en aquest sentit, contribueix al desenvolupament de l'avaluació psicològica proporcionant teories, mètodes i tècniques que, en darrer terme, permeten descriure, classificar, diagnosticar, explicar o predir els fenòmens psicològics objecte de mesura. De fet, contextualitzada així la psicometria podem anar un pas més enllà i assenyalar que, al seu torn, l'avaluació psicològica no es troba en cap altre lloc que al servei de la intervenció psicològica. És a dir, la finalitat última que guia el desenvolupament i l'administració de tests és la de contribuir a la recollida de les evidències necessàries que permetin als psicòlegs prendre una decisió o orientar alguna acció. Evidentment aquesta intervenció dependrà del context en què l'avaluació hagi estat desenvolupada, però és important tenir-ho present per a entendre no solament què és la psicometria, sinó la seva importància per a l'exercici professional dels psicòlegs en els diferents contextos en què intervenen.

Partint d'aquest enfocament comprensiu de la psicometria, podem sintetitzar les diferents aproximacions i formular finalment una definició que incorpori el seu paper en el marc de l'avaluació i la intervenció psicològiques.

Així, la psicometria és una branca de la psicologia que, mitjançant teories, mètodes i tècniques vinculats al desenvolupament i l'administració de tests, s'ocupa de la mesura indirecta dels fenòmens psicològics amb l'objectiu de fer descripcions, classificacions o diagnòstics, donar explicacions o fer prediccions que permetin orientar una acció o prendre decisions sobre el comportament de les persones en l'exercici professional de la psicologia.

Per concloure aquest exercici de definició formal és important discutir breument les diferències que hi ha entre la psicometria i una altra disciplina també interessada per la mesura psicològica com és la **psicologia matemàtica**⁷. En aquest sentit, podem reconèixer la psicologia matemàtica com la branca de la psicologia interessada pel desenvolupament de models dels processos perceptius, cognitius i motors amb l'objectiu d'establir lleis que relacionin estímuls físics amb els comportaments a partir d'estudis quasi exclusivament experimentals. Són molts els autors comuns a les dues orientacions des dels orígens de l'estudi sistemàtic de les diferències humanes al segle XIX però, a diferència de la psicometria, la psicologia matemàtica no està tan interessada per les diferències individuals com per la definició de lleis generals que modelitzin el comportament mitjà de les persones.

⁽⁷⁾Podeu veure, entre d'altres, Jáñez (1989), i Padilla, Merino, Rodríguez-Miñón i Moreno (1996).

2. Fonaments de la psicometria

Abordada la psicometria des d'una perspectiva històrica, tractant les seves arrels i la seva construcció com a disciplina científica en el context de la psicologia, en aquest segon apartat en revisarem els fonaments. Començarem discutint una definició i classificació dels tests, progressarem pels diferents models de mesura psicomètrica desenvolupats en les diferents teories dels tests, i farem una breu introducció a la teoria més estesa en la pràctica actual: la teoria clàssica dels tests. Finalment, recapitularem aquests fonaments atenent al procés que la psicometria segueix per establir les seves inferències sobre els fenòmens psicològics no observables a partir de les puntuacions obtingudes mitjançant tests. Comencem, doncs, amb una definició i classificació dels tests.

2.1. Definició i classificació dels tests

D'acord amb la definició del manual clàssic d'Anastasi (1988), un test psicològic⁽⁸⁾ és un procediment de mesura objectiva i estandarditzada d'una mostra de comportaments. Altres definicions són possibles⁽⁹⁾ però aquesta conté tres elements fonamentals que ens permeten abordar sistemàticament les característiques més importants que compleixen de manera genèrica els tests. D'una manera o altra, ja les hem anat introduint en aquest text.

⁽⁸⁾En endavant, en direm un test.

⁽⁹⁾Podeu veure, entre d'altres, Kaplan i Saccuzzo (2001), Murphy i Davidshofer (2005) o Urbina (2004).

En primer lloc, la mesura mitjançant el desenvolupament i l'administració de tests és, o al menys pretén ser, **objectiva**. En aquest sentit, l'objectivitat fa referència a la substitució del judici personal basat en criteris subjectius per un conjunt de normes determinades i conegudes que permeten obtenir i interpretar les puntuacions dels individus en igualtat de condicions. Així mateix, la mesura que cerquen els tests és **estandarditzada**, en tant que les puntuacions obtingudes depenen d'un procediment establert d'administració, correcció i interpretació que les fan, o les haurien de fer, invariants del professional que administra els tests, les condicions específiques en què ho fa, i la manera en què obté i interpreta les puntuacions resultants de la mesura. Finalment, els tests adrecen els fenòmens psicològics no observables mitjançant una **mostra de comportaments**. En la mesura que aquesta mostra sigui representativa del conjunt, les puntuacions obtingudes permetran als professionals establir adequadament les seves inferències sobre el comportament general de les persones més enllà dels elements específics avaluats mitjançant tests.

Ja hem discutit la naturalesa no observable dels fenòmens psicològics i la dificultat intrínseca que representa per a la **mesura indirecta mitjançant tests** que es proposa la psicometria. Però és important destacar aquesta qüestió quan la comparem amb altres tipus de mesura científica. En aquest sentit, la mesura dels fenòmens psicològics no resulta tan senzilla com la de les magnituds físi-

ques dels objectes –per exemple, el seu pes o la seva longitud–, que les ciències naturals poden observar i manipular directament. A més a més, tal com assenyala García Cueto (1993), hi ha d'altres **diferències importants** entre aquest tipus de mesura i la mesura indirecta dels fenòmens psicològics que han de ser tinguts en compte.

D'una banda, l'objectiu habitual de la mesura de les magnituds físiques és obtenir informació sobre un únic objecte. La psicometria, en canvi, és proposa desenvolupar instruments que permetin obtenir-la sobre un grup d'individus amb l'objectiu de treure conclusions sobre cadascun d'ells, sobre el grup sencer, i fins tot extrapolar-ne els resultats a les poblacions de referència d'on provenen aquests individus. D'altra banda, la mesura de magnituds físiques parteix de la possibilitat de repetir-ne el procediment tantes vegades com es vulgui sense variar les condicions en què es porta a terme la mesura. Aquest procediment és incompatible amb la mesura mitjançant tests, en tant que la repetició en l'aplicació d'una mateixa prova sobre els mateixos individus produeix variacions en les puntuacions que poden ser explicades pel cansament o per l'efecte de la pràctica en l'aprenentatge i no per variacions substancials en els fenòmens d'interès. Malgrat això, podem trobar algunes pràctiques desenvolupades en els contextos aplicats de les ciències naturals que tenen molt en comú amb l'administració de tests.

Així, quan es vol analitzar la contaminació atmosfèrica d'una ciutat o el nivell d'alcohol en sang d'una persona, els científics desenvolupen un conjunt de proves objectives –anomenades *reactius*– per tal d'aplicar-les de manera estandaritzada sobre una mostra representativa de l'objecte que volen mesurar –tant si és l'aire de la ciutat com la sang de l'individu, seguint els mateixos exemples. En aquest sentit, el procés de desenvolupament i administració de tests és comparable amb aquestes pràctiques, i posa en relleu el sentit que la psicometria atribueix als tests com els seus instruments de mesura. Un test, lluny del domini de la psicologia, no és més que una prova, un examen o un assaig. En definitiva, és un **reactiu** que, en el cas de la mesura indirecta dels fenòmens psicològics mitjançant tests, no té cap altra finalitat que produir una reacció en el comportament dels individus per a registrar-lo i obtenir una puntuació com a resultat del procés de mesura. Aquest sentit etimològic de la paraula *test* el podem trobar en l'arrel de la seva definició en el context de la psicometria (per exemple, vegeu Yela, 1984), en què s'assimila directament amb el reactiu de les ciències naturals i se n'espera que, aplicat a un individu, reveli i doni testimoni fidel dels fenòmens psicològics no observables que són objecte de mesura.

Des d'aquest punt de vista, el valor d'un test es troba en la capacitat per a **sus-citar i mesurar comportaments** que resultin un bon indicador –és a dir, una bona representació– del conjunt global de comportaments implicats en els fenòmens d'interès. En aquest sentit, hem de fer una precisió important amb relació als fenòmens que són objecte de mesura mitjançant tests. En aquest text parlem dels fenòmens psicològics d'una manera general i ho fem en tant

que l'objectiu d'un test és la mesura objectiva i estandarditzada d'una mostra de comportaments. Això no vol dir, però, que els tests hagin d'adreçar-se exclusivament a elements amb una llarga tradició en la teoria psicològica com són, entre d'altres, la intel·ligència o la personalitat. Per exemple, alguns tests s'adrecen a la mesura de les respostes fisiològiques, com poden ser la conductivitat de la pell o la freqüència cardíaca. D'altres es fixen en qüestions com les actituds racistes, el consum de substàncies o la sociabilitat de les persones. Finalment, en el camp de l'educació, no són pocs els tests que tenen per objectiu avaluar el domini dels estudiants en diferents àrees de coneixement. En tots el casos parlem de fenòmens psicològics en tant que responen a objectes d'interès per a les ciències socials i del comportament i, a més a més, recauen en els dominis de la mesura indirecta mitjançant tests sempre que es reuneixin les condicions que estableix la psicometria.

Així, la mesura indirecta que es proposen els tests comença amb una definició precisa dels fenòmens objecte d'interès i se sustenta fonamentalment en dos tipus de teories. D'una banda, en una **teoria substantiva** sòlida i ben establerta sobre el comportament de les persones –per exemple, una teoria de la intel·ligència–, que donarà el suport teòric necessari per a definir els elements crítics que conformen aquests fenòmens i determinarà els comportaments implicats que seran emprats com a evidències observables en el procés de mesura. De l'altra, en una **teoria dels test**, que permet establir les inferències sobre els fenòmens psicològics no observables a partir de les puntuacions obtingudes en el procés de mesura. Al seu cor, un **model de mesura** determinat que serveix al propòsit de relacionar les puntuacions obtingudes i els fenòmens objecte de mesura, a partir del qual s'articula el **procés d'inferència psicomètrica**. Tant els models de mesura com el mateix procés d'inferència psicomètrica seran tractats en detall més endavant. Però és important fer-ne esment en aquest punt perquè resulten clau per a entendre la mesura indirecta que es proposen els tests i que els defineix amb relació a d'altres tipus de mesura científica.

Finalment, una última consideració sobre els tests i els diferents termes amb què es fa referència amb ells ens porta a fixar-nos en dues grans formes de tractar les respostes obtingudes per mesurar els fenòmens psicològics d'interès. D'una banda, els anomenats **tests d'habilitat o de potència** tenen com a objectiu avaluar la competència, l'aptitud o els coneixements dels individus a partir de l'encert o la qualitat de les seves respostes. Són proves que discriminen respostes correctes i incorrectes i és aquesta la base per a puntuar les execucions individuals. De l'altra, els anomenats **tests de personalitat**¹⁰ tenen un objectiu diferent i cerquen conèixer de manera general les motivacions, preferències, opinions o actituds dels individus envers a un determinat estímul. Aquest segon tipus de tests no tenen respostes correctes i per tant no serveixen per a avaluar l'encert o l'error dels individus. Aquesta distinció és rellevant des d'un punt de vista terminològic, ja que els tests de personalitat són moltes vegades anomenats **qüestionaris, inventaris o escales**, malgrat que aquest ús no és sempre consistent. De fet, el terme *escala* pot esdevenir sinònim de *test* quan aquest està conformat per diferents parts, o fins i tot de cadascuna d'aquestes

(10) Tot i que el terme personalitat pot fer referència a la branca de la psicologia interessada per les característiques personals que influeixen les seves cognicions, emocions, motivacions i comportaments, en aquest cas s'utilitza en un sentit més ampli per a fer referència als tests que, en oposició als d'habilitat o de potència, no pretenen mesurar les capacitats de les persones.

parts per a reflectir les dimensions o característiques específiques que mesuren en el context del test global. D'altra banda, un altre terme estès també és el de **bateria**, que tampoc no té un significat unívoc i pot fer referència tant a un test compost per diverses parts com a una selecció de diferents tests administrats conjuntament per un professional en una avaluació psicològica determinada.

Després d'aquesta discussió terminològica, i malgrat l'enorme diversitat de tests que hi ha actualment, estem en disposició de classificar-los en funció d'algunes de les seves característiques més importants. Així:

- **Segons el propòsit.** El tests poden tenir diferents finalitats, i entre elles en podem destacar dues de fonamentals: la **diagnosi**, orientada a l'avaluació de les condicions actuals dels individus, i la **predicció**, que es proposa relacionar la mesura actual amb el comportament de les persones en situacions futures. En aquets sentit, per exemple, una cosa és obtenir informació sobre la destresa d'una persona en la resolució de problemes i una altra és utilitzar aquesta informació per a predir el seu desenvolupament en una feina determinada. Altres propòsits més específics són també possibles, com ja hem comentat en la discussió de la definició formal de la psicometria.
- **Segons el contingut.** De manera general, podem classificar els tests segons l'àrea de la psicologia a què pertanyen els fenòmens psicològics que volem mesurar. Aquesta classificació, però, varia segons els autors que la duen a terme, en tant que es poden fixar en diferents nivells de complexitat en jerarquitzar aquestes àrees. En el nivell més general podem distingir bàsicament tres grans grups que inclourien els tests orientats a l'avaluació de les **habilitats cognitives, la personalitat i les actituds**.
- **Segons el format.** Tenint en compte els materials utilitzats és freqüent distingir els tests de llapis i paper, de manipulació i de mesures fisiològiques. Ens referim als **tests de llapis i paper** quan presenten els ítems o preguntes en paper i requereixen de l'individu que doni algun tipus de resposta escrita. Els **tests computeritzats**, malgrat que no fan servir el paper, podrien ser encabits en aquesta categoria en tant que exigeixen algun tipus de resposta escrita mitjançant dispositius electrònics. D'altra banda, són **tests de manipulació** els que presenten una sèrie d'objectes, imatges o trencaclosques que els individus han de resoldre per a demostrar la seva habilitat. Finalment, els **tests de mesures fisiològiques** utilitzen sensors de diversa mena per a registrar les reaccions dels individus davant dels estímuls físics presentats.
- **Segons el tipus d'administració.** Una altra manera de classificar els tests pot tenir en compte la manera en què són utilitzats i ens permet distingir els **tests individuals i grupals**. Aquesta classificació se superposa a la que hem fet segons el format però ens permet distingir tests que requereixen una administració a un únic individu o permeten utilitzar-lo amb un grup

d'individus al mateix temps. També podem distingir els **tests verbals i no verbals**, en funció de si es fa una presentació oral o escrita del test i les seves instruccions o no.

- **Segons el tractament de les respostes.** Seguint la diferència que establíem entre els **tests d'habilitat o de potència** i els **tests de personalitat**, podem classificar també els tests en dos grans grups en funció de si avaluem l'encert en les respostes per a determinar la competència, l'aptitud o els coneixements dels individus, o si tracten d'avaluar motivacions, preferències, opinions o actituds. Aquests dos tipus de proves són també conegudes com a **proves d'execució màxima** i **proves d'execució típica**, respectivament.
- **Segons la interpretació de les puntuacions.** Tal com ja hem comentat, les puntuacions obtingudes mitjançant tests poden ser interpretades sobre la base de la norma o en referència a un criteri. Així, els **tests normatius** permeten comparar la puntuació de l'individu amb l'execució observada en un grup de referència que prèviament ha respost al mateix test. Aquest grup pot estar compost mitjançant mostreig probabilístic, quan representa la població de referència, o no probabilístic, quan estem interessats a comparar les puntuacions amb un grup de persones que compleix unes característiques determinades. D'altra banda, els **tests referits a criteri** prenen com a referència la definició d'un domini de coneixements o habilitats específics i permeten mesurar l'execució de l'individu, no ja en comparació amb un grup de referència, sinó en funció del seu grau d'adequació o consecució d'aquest criteri.
- **Segons l'estatus comercial.** Finalment, els tests es poden classificar en funció de si són propietaris o no. Els **tests propietaris o comercials** són tests que requereixen el pagament pel seu ús i inclouen els exemplars del mateix test i un manual que conté informació sobre el seu desenvolupament i les seves propietats psicomètriques, els fulls de correcció de les respostes i les taules per a interpretar les puntuacions amb relació als grups de referència. En d'altres casos parlem de **tests oberts o no comercials**, i generalment es poden obtenir contactant amb l'autor, el qual, habitualment, també ha desenvolupat alguna mena de manual que contextualitza el test. De vegades aquesta informació queda reduïda a la publicació d'un article científic en què l'autor l'introdueix, fa una primera administració i presenta les seves propietats psicomètriques.

2.2. Models de mesura psicomètrica

Qualsevol tipus de mesura científica, i la mesura indirecta dels fenòmens psicològics mitjançant tests no n'és una excepció, es fonamenta en una **definició del mateix concepte de mesura**. En aquest sentit, tal com ja hem introduït, en el nucli de les teories dels tests podem trobar un conjunt de models que

estableixen què és exactament mesurar i, com a conseqüència, especifiquen la relació entre les puntuacions empíriques obtingudes i els fenòmens psicològics no observables objecte d'interès. Són els **models o paradigmes de mesura**, que la psicometria ha desenvolupat de manera formal a partir del debat iniciat als anys quaranta sobre la viabilitat de la mesura dels fenòmens psicològics. Abans de caracteritzar aquests models, però, reprendrem la controvèrsia a partir de les conclusions de Ferguson i els seus col·laboradors (1940) per a la British Association for the Advancement of Science. Com veurem a continuació, representen un punt de partida important des del qual la psicometria ha desenvolupat els seus propis models de mesura per abordar específicament els fenòmens psicològics no observables.

D'acord amb el treball de Campbell (1920), físic i membre de la comissió que va elaborar aquest informe, el **model general de mesura desenvolupat en la física** es basaria en l'existència d'una equivalència –un isomorfisme, en les seves paraules– entre alguns atributs físics i les propietats additives dels números. Mesurar seria assignar números a aquests atributs per a representar l'additivitat dels objectes físics i aquesta representació seria només possible en tant que es satisfessin dues condicions: que l'atribut es pogués ordenar, és a dir, que fos possible determinar que l'atribut present en un objecte és menor, igual o més gran que en un altre; i que es pogués demostrar empíricament la presència de l'additivitat mitjançant la manipulació empírica. És el que va anomenar *mesura fonamental*, definint la mesura dels atributs físics per analogia amb els números a partir de les operacions portades a terme sobre els objectes per a quantificar-ne els atributs.

En un exemple senzill, el pes de dos o més objectes pot ser observat empíricament mitjançant una balança per a ordenar-los de més petit a més gran, com també es pot demostrar empíricament que el pes obtingut després de posar aquests objectes junts a la balança dóna com a resultat la suma dels pesos parcials.

Campbell (1928) va estendre aquest model general al que va anomenar *les mesures derivades*, en què la quantificació de certs atributs físics no dependria de la manipulació empírica sinó del descobriment de relacions matemàtiques entre dues mesures fonamentals. Per exemple, seria el cas de la densitat dels objectes, mesura derivada que respondria a una raó entre la seva massa i el seu volum. En qualsevol cas, la clau de la mesura per al model de Campbell es trobaria en la necessitat d'observar i manipular empíricament els objectes, bé per a quantificar un atribut físic mitjançant la seva observació directa o bé per a establir funcions matemàtiques que relacionin altres atributs observables. En la mesura en què els fenòmens psicològics no poden ser observats ni manipulats empíricament, ni són producte d'una relació entre d'altres atributs observables, la mesura indirecta que persegueix la psicometria no seria equiparable a la mesura fonamental o derivada de la física i, per tant, resultaria inviable. Aquesta va ser la conclusió final de la comissió organitzada per la British Association for the Advancement of Science, que, com ja hem comentat, lluny de tancar la controvèrsia va obrir un dels debats més importants per a la mesura indirecta dels fenòmens psicològics mitjançant tests.

Com a resposta a les restriccions imposades pel model general de mesura desenvolupat en la física, la psicometria va avançar en el debat sobre l'aproximació indirecta als fenòmens psicològics desenvolupant les seves pròpies teories i models de mesura. Tot i que altres autors utilitzen altres termes¹¹ en aquest text seguirem a Michell (1986) i els anomenarem *model representacional*, *operacional* i *clàssic*. A continuació, caracteritzarem aquests models i analitzarem breument les diferències en el seu plantejament del procés de mesura per tal de fer-lo extensiu als fenòmens psicològics. Essent un debat encara obert, i susceptible de diferents interpretacions segons els diferents autors, optem per l'exposició de Michell en tant que planteja, des d'una visió crítica, les incògnites més importants que la psicometria ha de resoldre en la seva construcció com a disciplina científica.

En primer lloc, el **model representacional** va ser desenvolupat a partir dels treballs de Stevens (1946) i Suppes (1951), i defineix la mesura com el procés d'assignació de números a objectes a partir d'unes regles de manera que reflectixin relacions empíriques entre els objectes. Aquestes relacions, de manera comparable al model de Campbell, quedarien representades per les propietats dels números però no se centrarien exclusivament en l'additivitat. De fet, Stevens va obrir la porta a altres tipus de relacions per a portar a terme la mesura, acceptant per exemple l'equivalència o l'ordre. Mesurar no seria únicament quantificar sinó la representació numèrica dels fets empírics en sentit ample, i donaria com a resultat quatre escales clàssiques per a la psicologia segons el tipus de relació representada: escala nominal, ordinal, d'interval i de raó.

Tal com fa el model desenvolupat en la física, el model representacional assumeix la correspondència entre les propietats dels números –les puntuacions obtingudes mitjançant tests, en aquest cas– i les relacions dels objectes que representen. En canvi, trasllada el focus des de la demostració empírica d'aquestes relacions a l'establiment d'un conjunt de restriccions sobre el tipus d'operacions estadístiques permeses a cada escala que garanteixi la invariància de les relacions empíriques representades. És el debat sobre l'estadística permesa, que ha portat a una gran controvèrsia entre els partidaris del model representacional i l'operacional.

En segon lloc, el **model operacional** es basa en les contribucions de Bridgman (1927) i Dingle (1950), a partir de les quals es va proposar la definició de qualsevol concepte mitjançant les operacions necessàries per a mesurar-lo. Així, d'acord amb aquest model, la mesura no seria una altra cosa que el conjunt d'operacions necessàries per a definir un concepte que, en darrer terme, acaben produint números. Aquesta definició de la mesura té punts en comú amb la del model representacional, ja que assignar números –les puntuacions obtingudes mitjançant tests– d'acord amb unes regles és un tipus particular de les operacions possibles, i els números resultants no són altra cosa que el pro-

⁽¹¹⁾Per exemple, podeu veure Fraser (1980), Swistak (1990) o Hand (1996).

Per saber-ne més

L'estudiant interessat pot trobar una discussió molt més àmplia de la que podem plantejar en aquest text a Michell (1999).

Per saber-ne més

En relació amb el debat sobre l'estadística permesa, podeu veure, entre d'altres, la interessant discussió de Gaito (1980), i Velleman i Wilkinson (1993).

ducte de les operacions portades a terme. En canvi, una diferència important entre aquests dos models de mesura es troba en el requeriment segons el qual els números representen o no un sistema de relacions empíriques.

Per al model representacional, aquestes relacions empíriques són prèvies a la mesura i tenen una existència independent de les operacions portades a terme per a produir els números. El model operacional, en canvi, limita el domini de les operacions a les que d'una manera consistent produeixen números i, en cap cas, està interessat per l'existència d'una realitat empírica que doni suport a aquests números. Aquesta diferència en l'enfocament de la mesura, a més, té implicacions importants quant al debat sobre l'estadística permesa, ja que els partidaris del model operacional no veuen la necessitat d'establir cap restricció sobre les operacions estadístiques. En no ser l'objectiu la cerca de relacions empíriques comparables a les relacions entre els números, sinó els números en sí mateixos, cap restricció no pot ser aplicable a unes puntuacions que, en paraules de Lord (1953), no tenen consciència d'on provenen.

Finalment, el **model clàssic** es proposaria resoldre aquesta discussió plantejant la naturalesa quantitativa dels fenòmens psicològics com a condició per a la mesura. Partint dels treballs de Rozeboom (1966) i Jones (1971), i arrelat en el desenvolupament de les teories quantitatives de la psicologia de les dècades anteriors, aquest model defineix la mesura com la determinació de la quantitat –del quant– en què un atribut és present en l'objecte mesurat. És a dir, mesurar consisteix a determinar quantes unitats són presents en l'atribut observat però, a diferència dels models de Campbell i Stevens, no exigeix que hi hagi una relació empírica entre els objectes. De fet, el model clàssic mesura atributs, no objectes, i per tant no preveu en el seu nucli la demostració de l'additivitat mitjançant l'observació i la manipulació empírica dels objectes físics. Encara més, a diferència dels models representacional i operacional, la mesura no es considera l'assignació de números a objectes d'acord amb unes regles ni el resultat d'un tipus particular d'operacions que produeix números, respectivament, sinó que es defineix com el procés de descobriment de relacions numèriques entre els valors observats en un atribut quantitatiu.

Seguint l'exemple del pes de dos o més objectes físics, les seves mesures en una unitat determinada (per exemple, en quilos) proporcionen un sistema per a relacionar-los matemàticament en termes de les seves magnituds relatives, i permeten afirmar que un d'ells pesa n vegades més o menys que un altre. Aquestes relacions numèriques són, per al model clàssic, tan vàlides com ho són les relacions observables entre els objectes i per tant comparteixen el mateix estatus com a evidències empíriques necessàries per al desenvolupament científic. En aquest sentit, és molt important assenyalar el paper de la teoria substantiva, ja que és la base sobre la qual es fonamenten tant la hipòtesi sobre la naturalesa quantitativa dels fenòmens psicològics objecte de mesura com el procés de determinació de les relacions numèriques amb les mesures obtingudes. Finalment, el model clàssic no preveu les diferents escales de mesura acceptades pels models representacional i operacional. D'aquesta manera les

escales nominals i ordinals queden fora del domini de la mesura –que, recordem, és quantitativa per definició– i per tant no és necessària cap prescripció entorn del debat sobre l'estadística permesa.

Com ja hem assenyalat, els models de mesura psicomètrica es troben a la base de les diferents teories dels tests i durant les darreres dècades han proporcionat diverses alternatives per a l'aproximació indirecta a la mesura dels fenòmens psicològics que es proposa la psicometria. Així, aquests tres models de mesura –dèiem, representacional, operacional i clàssic– es corresponen amb la **teoria representacional**, la **teoria clàssica dels tests** i els **models de variable latent**, respectivament. Abordar totes aquestes teories alternatives està fora dels objectius d'un text introductori com aquest, tot i que l'estudiant interessat pot trobar una anàlisi molt detallada per exemple a Borsboom (2005). En canvi, optarem per restringir el tractament de les teories dels tests a la que gaudeix del suport majoritari en la pràctica actual de la psicometria. Parlem de la teoria clàssica dels tests (TCT), que és, a més a més, la base per al desenvolupament dels mòduls posteriors d'aquest text. És important destacar, per a evitar confusions, que la TCT no s'anomena així pel model de mesura psicomètrica que utilitza. De fet, la seva definició de mesura encaixa amb el model operacional i, en canvi, adopta aquest nom per oposició a d'altres enfocaments més moderns com la teoria de resposta a l'ítem (TRI). Finalment, introduïrem molt breument algunes de les innovacions que, d'acord amb el model clàssic de mesura propi dels models de variable latent, la TRI ha desenvolupat per tractar de resoldre algunes de les limitacions de la TCT.

2.3. Teoria clàssica dels tests

La TCT és la teoria dels tests més estesa actualment en la pràctica de la psicometria i es basa en el **model lineal clàssic** proposat per Spearman¹², sistematitzat per Gulliksen (1950) i reformulat posteriorment per Lord i Novick (1968), que articula el procés de mesura definint tres conceptes fonamentals: la puntuació vertadera, la puntuació empírica, i l'error de mesura. Per a introduir-lo, en aquest text seguirem l'exposició de Muñiz (1996 i 2003).

Així, partint del model operacional de mesura psicomètrica, aquesta teoria dels tests no està interessada pel sistema de relacions empíriques sinó que centra la seva atenció en l'anàlisi de les puntuacions obtingudes per tal d'estimar els errors comesos en el procés de mesura indirecta dels fenòmens psicològics. És l'anomenada **puntuació empírica (X)**, que, d'acord amb aquesta teoria, respondria a una relació lineal de dos components fonamentals:

$$X = V + e$$

⁽¹²⁾Per a una discussió històrica, podeu veure Traub (1997).

Per saber-ne més

L'estudiant interessat pot trobar explicacions més detallades a Borsboom (2005), Crocker i Algina (2006) o De Gruijter i Van der Kamp (2008).

D'una banda, la **puntuació vertadera (V)**, que seria el resultat ideal o desitjat en què el procés de mesura mitjançant tests hauria estat portat a terme lliure de qualsevol tipus d'error. De l'altra, l'**error de mesura (e)**, que seria responsable de la discrepància entre la puntuació vertadera que es vol aconseguir i la puntuació empírica obtinguda com a resultat de l'administració del test.

Com en qualsevol tipus de mesura científica, la mesura indirecta dels fenòmens psicològics es troba subjecte a aquestes variacions o desviacions no desitjades, i el model lineal clàssic proposa un conjunt de supòsits, que permeten fer una estimació de les puntuacions vertaderes a partir de les puntuacions empíriques obtingudes, i una definició de tests paral·lels. D'aquesta manera:

- **Primer supòsit:** $V = E(X)$. La puntuació vertadera (V) es defineix matemàticament com l'esperança matemàtica de la puntuació empírica (X). És a dir, el primer supòsit del model lineal clàssic assumeix que, en cas que fos possible l'administració d'un test un nombre infinit de vegades, la mitjana de les puntuacions empíriques obtingudes ens donaria com a resultat la puntuació vertadera del subjecte en el test. La puntuació empírica, per tant, no és un substitut de la puntuació vertadera sinó la millor aproximació disponible en un procés de mesura que, de manera explícita, és reconegut com a no lliure d'error.
- **Segon supòsit:** $\rho(v, e) = 0$. No hi ha correlació entre les puntuacions vertaderes dels subjectes (v) en un test i els errors respectius de mesura (e). És a dir, d'acord amb el segon supòsit, no s'espera que la grandària dels errors comesos estigui sistemàticament associada a la grandària de les puntuacions vertaderes.
- **Tercer supòsit:** $\rho(e_j, e_k) = 0$. Si disposem de dos tests diferents (j i k), no hi ha correlació entre els errors de mesura comesos amb cadascun d'ells (e_j i e_k , respectivament). És a dir, d'acord amb el tercer supòsit, els errors de mesura dels diferents tests són aleatoris cada vegada i, per tant, no s'espera que hi hagi cap relació entre ells.
- **Definició de tests paral·lels:** $V_j = V_k$ i $\sigma^2(e_j) = \sigma^2(e_k)$. Finalment, dos tests (j i k) són paral·lels sempre que les seves puntuacions vertaderes (V_j i V_k) i les seves variàncies dels errors de mesura ($\sigma^2(e_j)$ i $\sigma^2(e_k)$, respectivament), siguin idèntiques.

D'aquest model, amb els seus supòsits i la definició de tests paral·lels, es deriven un conjunt de deduccions immediates importants que formen la base per al mòdul dedicat a la fiabilitat en aquest text. Així:

- $e = X - V$. D'acord amb la formulació inicial del model lineal clàssic, l'error de mesura seria la diferència entre la puntuació empírica i la puntuació vertadera.
- $E(e) = 0$. L'esperança matemàtica dels errors de mesura és 0, de manera que, si fos possible administrar un test un nombre infinit de vegades, aquests errors aleatoris o no esbiaixats es compensarien o anul·larien entre ells.
- $\mu_x = \mu_v$. La mitjana de les puntuacions empíriques (μ_x) és igual a la mitjana de les puntuacions vertaderes (μ_v).
- $\text{cov}(V, e) = 0$. D'acord amb el segon supòsit del model, les puntuacions vertaderes no covarien amb els errors de mesura.
- $\text{cov}(X, V) = \text{var}(V)$. La covariància entre les puntuacions empíriques i vertaderes és igual a la variància de les puntuacions vertaderes.
- $\text{cov}(X_j, X_k) = \text{cov}(V_j, V_k)$. La covariància entre les puntuacions empíriques de dos tests (X_j i X_k) és igual a la covariància de les seves puntuacions vertaderes (V_j i V_k).
- $\text{var}(X) = \text{var}(V) + \text{var}(e)$. La variància de les puntuacions empíriques és el producte de la suma de la variància de les puntuacions vertaderes i la dels errors de mesura.
- $\rho(X, e) = \sigma_e / \sigma_x$. La correlació entre les puntuacions empíriques i els errors de mesura és el resultat de la divisió de la desviació típica dels errors entre la de les puntuacions empíriques.
- $\mu_1 = \mu_2 = \dots = \mu_k$. Per a k tests paral·lels, les mitjanes són idèntiques.
- $\sigma^2(X_1) = \sigma^2(X_2) = \dots = \sigma^2(X_k)$. De la mateixa manera, les variàncies de k tests paral·lels són idèntiques.
- $\rho(X_1, X_2) = \rho(X_1, X_3) = \dots = \rho(X_j, X_k)$. Finalment, les correlacions entre k tests paral·lels són també idèntiques.

Tenint en compte que ni les puntuacions vertaderes ni els errors de mesura són observables directament a partir de les puntuacions empíriques, cap dels supòsits ni les vuit primeres deduccions immediates són demostrables.

Com ja hem discutit anteriorment, la TCT resulta una aproximació útil per a l'estimació de les puntuacions vertaderes a partir de les puntuacions empíriques, però no està exempta d'**algunes limitacions** que, des del model lineal clàssic, no són fàcils de resoldre. Tal com van discutir Lord i Novick (1968) en la seva reformulació, aquesta teoria té dues dependències importants amb re-

Per saber-ne més

L'estudiant interessat pot trobar el desenvolupament matemàtic que sustenta aquestes deduccions a Muñiz (2003).

lació als instruments i les mostres utilitzades. D'una banda, una dependència de les puntuacions empíriques obtingudes respecte dels instruments, de manera que, per exemple, dos tests d'intel·ligència independents utilitzen escales diferents i, per tant, les puntuacions que se n'obtenen no resulten comparables directament. De l'altra, una dependència de les propietats psicomètriques dels tests respecte de les mostres utilitzades per a desenvolupar-los, de manera que, per exemple, la dificultat dels ítems que conformen un test depèn de les característiques dels individus a qui s'administra el test.

Tractant de resoldre aquests problemes, la TRI es proposa canviar el focus del tractament del test sencer al tractament individual dels ítems. Per fer-ho, aquesta teoria dels tests planteja l'existència d'una relació entre les puntuacions en la variable latent –recordem, d'acord amb el model clàssic de mesura psicomètrica– i la probabilitat d'encertar cada ítem introduint diferents funcions matemàtiques per a modelitzar adequadament aquesta relació. Així, una vegada seleccionada la funció més adequada es construeixen el que s'anomenen *les corbes característiques dels ítems*, una modelització de les respostes que pot tenir en compte diferents paràmetres com la capacitat de discriminació dels ítems, la seva dificultat o la probabilitat que siguin encertats a l'atzar.

No és aquest el lloc per a fer una discussió en profunditat, però sí per a fer èmfasi en les implicacions que tenen els diferents models de mesura psicomètrica a partir dels quals la TCT i la TRI es proposen la mesura indirecta mitjançant la utilització de tests. D'una banda, tal com fa la TCT, definint les puntuacions empíriques com la suma de les puntuacions vertaderes i l'error de mesura i, de l'altra, tal com fa la TRI, establint diferents funcions matemàtiques segons el model de relació entre la variable latent i la capacitat de discriminació dels ítems.

2.4. El procés d'inferència psicomètrica

Per concloure amb els seus fonaments, recapitularem el procés que la psicometria segueix per establir-ne les inferències sobre els fenòmens psicològics no observables a partir de les puntuacions obtingudes mitjançant tests. Abans d'abordar-ne les especificitats és important tenir present que la mesura objectiva i estandarditzada d'una mostra de comportaments s'ajusta, de manera general, al procediment establert pel **mètode científic**. El podríem resumir així:

- 1) Formular una pregunta d'investigació o una hipòtesi rellevants.
- 2) Especificar i definir adequadament totes les variables involucrades.
- 3) Desenvolupar o escollir els instruments i procediments necessaris per a portar a terme les mesures.

Per saber-ne més

L'estudiant interessat pot trobar una comparació molt entenedora d'aquestes dues teories dels tests a Muñiz (2010), com també desenvolupaments més detallats de la TRI a Muñiz i Hambleton (1997), Maydeu-Olivares i McArdle (2005), Borsboom (2005), o Martínez Arias, Hernández Lloreda i Hernández Lloreda (2006).

Una referència clàssica sobre els fonaments de la TRI és la de Hambleton, Swaminathan i Rogers (1991).

4) Avaluar el funcionament dels instruments i procediments per a obtenir les garanties suficients sobre la qualitat del procés de mesura.

5) Recollir les evidències necessàries que permetin respondre als objectius de la investigació.

6) Resumir i, sempre que sigui possible, tractar estadísticament les dades obtingudes per a determinar fins a quin punt els resultats són significatius i, per tant, no són producte de l'atzar.

D'acord amb aquest procediment general, la psicometria ofereix un conjunt de teories, mètodes i tècniques vinculats al desenvolupament i l'administració de tests que donen el suport necessari als punts 2, 3 i 4 del mètode científic quan s'utilitza en el context de la mesura indirecta dels fenòmens psicològics no observables. I ho fa, com hem assenyalat anteriorment, partint d'una definició precisa dels fenòmens objecte de mesura i d'una selecció dels comportaments implicats que seran emprats com a evidències observables. Tot i això, és important fer èmfasi en algunes dificultats específiques que, com a disciplina científica, la psicometria ha d'enfrontar per tal de garantir la confiança en les inferències establertes sobre els fenòmens psicològics no observables a partir de les puntuacions obtingudes mitjançant tests. Totes elles, en íntima relació amb el repte essencial de la psicometria, són conseqüència directa de la dificultat afegida que representa la impossibilitat d'observar i manipular directament els fenòmens psicològics objecte d'interès. En aquest sentit, entre els **reptes específics** en podem assenyalat alguns dels més importants:

- **La mesura indirecta dels fenòmens psicològics no és unívoca.** Com hem anat veient, no hi ha una única manera de definir el procés de mesura així com tampoc no hi ha una única manera d'entendre els propis fenòmens objecte de mesura. Són moltes les decisions que els professionals interessats pel desenvolupament i l'administració de tests han de prendre durant aquest procés i, per tant, són moltes les possibles solucions als problemes que sorgeixen en l'aproximació indirecta a la mesura dels fenòmens psicològics. Entre aquestes decisions destaquen les relacionades amb l'elecció del que és rellevant mesurar, la selecció i el mostreig dels comportaments observables vinculats i la definició de les característiques que els tests han de tenir per a suscitar-los i mesurar-los adequadament.
- **La teoria té un paper fonamental en la mesura psicomètrica.** Totes les decisions que prenen els professionals interessats pel desenvolupament i l'administració de tests han d'estar fonamentades, com hem assenyalat anteriorment, per dos tipus de teories. D'una banda, per una teoria substantiva ben establerta entorn dels fenòmens psicològics i els comportaments de les persones que servirà de context de referència per al procés de mesura. De l'altra, una teoria dels tests, que permetrà adoptar l'enfocament més adequat per a portar a terme la mesura i establir la relació entre les puntuacions obtingudes i els fenòmens psicològics no observables objecte

d'interès. Si els tests tenen sentit i serveixen com a instruments de mesura al camp de la psicologia és, precisament, gràcies al suport que aquests dos tipus de teories proporcionen durant tot el procés.

- **La selecció i el mostreig dels comportaments imposa limitacions importants.** Un tercer repte important per a la mesura indirecta que es proposa la psicometria és la selecció i el mostreig dels comportaments emprats com a evidències observables. En consonància amb la teoria substantiva, el desenvolupament i l'administració de tests parteix d'una definició que orienta cap a uns comportaments vinculats i no cap a d'altres que, a la vegada, han de ser mostrejats adequadament atesa la impossibilitat material de suscitar i mesurar el domini sencer de comportaments al qual pertanyen. De fet, un comportament determinat pot ser emprat com a evidència de fenòmens psicològics diferents en funció de la perspectiva teòrica amb què es fonamenta el procés de construcció dels tests. Limitacions importants vinculades al temps requerit i les condicions d'administració també han de ser resoltes adequadament per a fer els tests operatius des d'un punt de vista pràctic.
- **Les puntuacions obtingudes requereixen una interpretació adequada.** Les puntuacions obtingudes mitjançant la utilització de tests no són, *per se*, informatives i han de ser sempre interpretades per a respondre al propòsit amb què els tests han estat desenvolupats. Tant des d'un punt de vista general, amb relació al contingut i el seu propòsit com amb relació al sistema de referència emprat –recordem, d'acord amb la norma, és a dir, amb relació a l'execució d'un grup de referència, o referits a un criteri, és a dir, amb relació al grau d'adequació o consecució d'aquest criteri– la utilització de tests requereix uns coneixements i unes destreses que només podem trobar a mans de professionals qualificats i sempre ha de complir amb els objectius més generals que persegueix l'avaluació psicològica a la qual serveix.
- **La mesura dels fenòmens psicològics no està lliure d'error.** L'error és un component inherent a qualsevol procés de mesura i ha de ser objecte d'un tractament adequat per a conèixer-lo i minimitzar-lo. La mesura indirecta afegeix una dificultat específica que, de manera sintètica, trobem en la impossibilitat d'observar directament tant els fenòmens psicològics objecte d'interès com l'error comès durant el procés de mesura. Partint de les puntuacions obtingudes, i gràcies al suport dels models de mesura desenvolupats en les diferents teories dels tests, la psicometria dedica una bona part dels seus esforços a aconseguir la precisió necessària que, des d'un punt de vista científic, també és exigible a la mesura indirecta dels fenòmens psicològics.
- **La mesura psicològica no es pot entendre si no és amb relació a altres mesures o esdeveniments observables.** Tot i la importància de la teoria substantiva en el desenvolupament dels tests, no podem obviar final-

ment que la mesura indirecta dels fenòmens psicològics no té sentit únicament en si mateixa sinó també amb relació a d'altres mesures derivades de la mateixa teoria o d'altres comportaments observables que corroborin el procés de mesura mitjançant tests. Com hem dit anteriorment, la mesura mitjançant tests ha de proporcionar informació rellevant amb l'objectiu de fer descripcions, classificacions, diagnòstics, explicacions o prediccions que permetin orientar una acció o la presa d'una decisió sobre el comportament de les persones, i és per això que les puntuacions obtingudes mitjançant tests requereixen un suport extern a la seva pròpia formulació per a demostrar-ne la utilitat pràctica.

Per a enfrontar-se a aquests reptes específics, la mesura indirecta dels fenòmens psicològics es proveeix d'uns **principis bàsics** que garanteixen la confiança en les inferències establertes a partir de les puntuacions obtingudes mitjançant tests. Aquests principis serveixen per a avaluar les propietats psicomètriques dels ítems i dels tests en conjunt, i entre aquests principis bàsics destaquen els següents:

- **Fiabilitat:** la precisió amb què els tests porten a terme la mesura.
- **Validesa:** la confiança en el fet que les mesures es corresponen realment amb el que es proposen mesurar.

Els diferents mètodes i procediments vinculats a aquests principis són la clau de l'èxit dels tests com a instruments d'avaluació psicològica i seran, per tant, objecte d'un tractament en profunditat en mòduls posteriors.

3. Construcció i administració de tests

Presentats els fonaments de la psicometria, en aquest darrer apartat ens farem càrrec del procés de construcció i administració de tests. Per fer-ho, començarem desenvolupant les diferents fases implicades en el disseny i construcció d'un nou test. A continuació, i partint de les claus que ens proporciona el coneixement d'aquest procés de construcció, discutirem alguns criteris importants per tal d'avaluar les característiques dels tests disponibles en la literatura i valorar-ne la conveniència amb relació als objectius de l'avaluació psicològica que han de servir. Finalment, abordarem els aspectes ètics i deontològics vinculats a la utilització de tests en el context general de la pràctica professional de la psicologia.

3.1. El procés de construcció de tests

El desenvolupament d'instruments de mesura és un procés fonamental per a qualsevol disciplina científica. I no ho és menys per a la psicometria, en què com ja hem discutit abastament la impossibilitat d'observar i manipular els fenòmens objecte d'interès afegeix una complexitat important. A continuació abordarem **deu fases fonamentals** en què, a tall de conclusió, podem organitzar el procés de construcció d'un nou test. Aquestes fases ens permeten donar una perspectiva de conjunt a aquesta introducció a la psicometria i fer una síntesi dels conceptes més importants que hem anat discutint.

Tot i que no són necessàriament seqüencials, i en tot cas no els cal seguir estrictament aquest ordre, aquestes deu fases representen algunes de les decisions més importants que els professionals interessats pel desenvolupament de nous tests han de prendre. Com veurem més endavant, conèixer aquest procés és fonamental no solament per a garantir la qualitat en el disseny i construcció de nous tests, sinó també perquè ens permet desenvolupar alguns criteris importants per a l'avaluació dels tests disponibles en la literatura. Així:

1) Delimitació de la finalitat del test. El desenvolupament d'un nou test comença amb una determinació clara del propòsit per al qual es vol recollir informació rellevant en el context de l'avaluació psicològica. Tal com hem assenyalat, els tests poden servir multitud de finalitats però entre elles en podem destacar algunes d'importantes com són descriure, classificar, diagnosticar i explicar el comportament de les persones o fer prediccions sobre aquest. Un propòsit ben delimitat és la primera condició a complir per a l'èxit en la construcció del nou test i en permetrà l'encaix en el procés general d'avaluació psicològica.

Per saber-ne més

L'estudiant interessat pot trobar explicacions més detallades a Muñiz (1996), Muñiz i Fonseca-Pedrero (2008), Murphy i Davidshofer (2005), Downing (2006), Chadha (2009) o Rust i Gollombok (2009).

2) Definició dels fenòmens psicològics objecte de mesura. D'acord amb el procés d'inferència psicomètrica, el segon pas per a la construcció d'un test consisteix en la delimitació precisa dels fenòmens que es volen mesurar. Per a fer-ho, recordem-ho, és necessària una teoria substantiva sòlida i ben contrastada entorn dels fenòmens psicològics i els comportaments de les persones que servirà de referència en el procés de la mesura mitjançant el nou test. Aquest marc teòric és fonamental en el moment de planificar una representació adequada dels fenòmens psicològics, i permet conèixer en detall altres instruments desenvolupats prèviament en la investigació psicològica. Avaluar-ne les limitacions i els punts forts és una bona guia per a enfocar el treball necessari per a donar suport teòric al nou test.

3) Selecció i mostreig dels comportaments observables. Establert el fenomen objecte de mesura, la teoria substantiva proporciona també el context necessari per a escollir els comportaments implicats que seran emprats com a evidències observables. La seva representació adequada és fonamental per a no ometre cap comportament rellevant i també per a evitar incloure'n d'altres no directament relacionats amb l'objecte de mesura. A vegades també és recomanable portar a terme observacions, entrevistes a informadors clau o grups de discussió que proporcionin informació complementària sobre els comportaments d'interès a partir de les experiències dels participants. Com ja hem discutit, aquesta no és una fase menys importat, en tant que és la base per al desenvolupament correcte dels ítems que conformaran el nou test.

4) Especificació de les característiques del test. Delimitat l'objectiu, definit l'objecte de mesura i seleccionades les evidències observables necessàries, el pas següent consisteix a triar les característiques del nou test per a suscitar i mesurar adequadament els comportaments d'interès. En primer lloc, decidirem si es tracta d'un test d'habilitat o de potència, en què s'avaluaran l'encert i l'error en les respostes per tal de determinar la competència, l'aptitud o els coneixements dels individus, o d'una prova per a avaluar motivacions, preferències, opinions o actituds. És a dir, si es tracta d'una prova d'execució màxima o d'execució típica, com també són conegudes. D'acord amb la classificació dels tests que hem fet en aquesta introducció a la psicometria, també és el moment per a decidir el format del nou test optant per una prova de llapis i paper, de manipulació o de mesures fisiològiques. Així mateix s'han de preparar els materials necessaris i es determinarà el tipus d'administració més adequada, triant bàsicament entre una prova individual o d'administració en grup. Finalment, també es decidirà el mètode més adequat per a interpretar les puntuacions obtingudes, a partir de la norma o amb referència a un criteri.

5) Desenvolupament dels ítems que conformaran el test. Especificades les característiques generals del nou test, en aquesta fase es portarà a terme el desenvolupament dels elements que contindrà. Aquesta no és una tasca senzilla i generalment comporta la col·laboració d'un grup d'experts en el camp per a trobar la millor representació de les mostres de comportament seleccionades. De fet, no és poc habitual desenvolupar molts més ítems dels estricta-

ment necessaris per a avaluar-ne el comportament en el marc del test i seleccionar els més adequats amb relació al propòsit del nou test. L'objectiu final és disposar de tants com siguin necessaris per a representar adequadament les diferents dimensions dels fenòmens psicològics objecte d'interès. Aquest també és el moment de decidir el format que adoptaran les respostes, que serviran per a codificar els comportaments de manera estructurada i d'acord amb unes regles clares.

6) Elecció d'una teoria dels tests. Tancant el cercle establert en el procés d'inferència psicomètrica, la fase següent en el procés de construcció implica la tria d'una teoria dels tests que, mitjançant un model de mesura psicomètrica determinat, servirà per a relacionar els fenòmens psicològics objecte d'interès i les puntuacions obtingudes mitjançant els ítems que conformen el nou test. Aquesta elecció és cabdal tenint en compte les seves conseqüències en els mètodes i tècniques emprats posteriorment per a avaluar les propietats dels ítems i del test en conjunt amb l'objectiu de garantir la confiança en les inferències establertes sobre el comportament de les persones a partir de les puntuacions obtingudes.

7) Realització d'una prova pilot. Una vegada construït el test, incloent-hi els ítems potencials i el format de les respostes, és necessari redactar les instruccions que l'acompanyaran i definir les condicions en què serà administrat. Una prova pilot servirà per a avaluar el grau de comprensió d'aquestes instruccions i la viabilitat de les condicions per a administrar el test, detectar possibles dificultats en el moment de registrar les respostes i fer una primera anàlisi de les propietats, tant dels ítems com del test en el seu conjunt, d'acord amb els mètodes i tècniques indicats per la teoria dels tests utilitzada. És el moment de valorar-ne el comportament en el procés de mesura i, a partir d'aquesta informació, refinar el test modificant, descartant o afegint nous ítems. Es poden requerir proves addicionals per a avaluar correctament les modificacions introduïdes abans de tancar aquesta fase i procedir amb el desenvolupament de l'estudi de camp final.

8) Desenvolupament de l'estudi de camp. Una vegada establert el test definitiu, el pas següent consisteix a administrar-lo a la població a què es dirigeix. Per a fer-ho, se selecciona la mostra de participants necessaris que, com ja hem assenyalat, pot ser probabilística o no probabilística en funció de si volem representar la població de referència o avaluar un grup de persones que compleix unes determinades característiques. Amb aquesta informació es desenvolupen les normes o barems per a permetre la interpretació de les puntuacions amb relació a l'execució del grup de referència. En cas que es tracti d'un test referit a criteri, en lloc de normes o barems es determinen els punts de tall que permetran distingir els diferents graus d'adequació o consecució del criteri. Així mateix, s'aprofundirà en el treball sobre les propietats psicomètriques dels ítems i del test en el seu conjunt, atenent especialment a la fiabilitat i validesa de les

mesures obtingudes. Aquesta és una part fonamental del procés de construcció de tests i és la que en darrer terme garanteix que la mesura psicomètrica compleix amb totes les exigències científiques.

9) Elaboració del manual del test. Amb tota aquesta informació es desenvolupa la documentació que acompanyarà al nou test, en què s'ha d'incloure informació rellevant relativa a les diferents fases involucrades en la seva construcció: fonamentació teòrica, finalitat i població a la qual s'adreça, instruccions per a l'administració, informació per a la interpretació de les puntuacions obtingudes i anàlisi de les propietats psicomètriques. Les primeres publicacions a revistes científiques serveixen per a començar a difondre tota aquesta informació, que com ja hem comentat no sempre acaba constituint un manual pròpiament dit.

10) Revisió i millora del test. Amb la publicació del test, ja sigui difonent-lo sota llicència o lliurement, el test es posa a disposició de la comunitat científica per a obtenir noves evidències que, mitjançant el treball independent de diferents investigadors, serviran per a millorar el coneixement sobre el seu funcionament i les seves propietats psicomètriques, així com per a adaptar-lo a altres entorns socioculturals o a d'altres poblacions diferents per a les quals ha estat desenvolupat el nou test. Canvis en els fenòmens psicològics objecte de mesura, nous avenços en l'aproximació teòrica a aquests fenòmens o l'adaptació del test a noves condicions d'administració o poblacions d'interès són algunes de les raons que justifiquen un treball de revisió i actualització pràcticament indefinida per tal de refinar-ne el funcionament i valorar-ne adequadament la utilitat al servei dels objectius de l'avaluació psicològica.

3.2. Criteris per a la valoració de tests

Construir un test nou no és la pràctica més habitual en l'exercici professional de l'avaluació psicològica. De fet, com hem vist, és un treball laboriós que requereix un esforç important per part de diversos professionals durant un període de temps prolongat per a planificar, executar i analitzar les dades recollides en un o més estudis de camp. És per això que la majoria de les situacions en què s'utilitzen els tests no comencen necessàriament amb la creació d'un original. En canvi, parteixen de la cerca i valoració d'algun test existent en la literatura que encaixi amb els objectius de l'avaluació psicològica. Per tal de simplificar aquest procés, organitzacions vinculades a l'exercici professional de la psicologia com l'American Psychological Association (<http://www.apa.org>) o el Consejo General de Colegios Oficiales de Psicólogos (<http://www.cop.es>) suggereixen algunes fonts per a la cerca de tests:

- **Publicacions periòdiques.** Amb format de llibre o d'enciclopèdia, diverses publicacions es dediquen a recopilar els tests comercials disponibles amb diferents nivells de detall. Per exemple, el Buros Center for Testing de la Universitat de Nebraska edita *Tests in Print* (TIP, <http://buros.org/tests-print>) i *Mental Measurements Yearbook* (MMY, <http://buros.org/mental-me>

asurements-yearbook). TIP és una de les recopilacions en llengua anglesa més exhaustives que proporciona informació rellevant sobre els objectius, les poblacions a qui s'adrecen, els autors i editorials que els publiquen i el preu de més de 3.000 tests. Per la seva banda, MMY va més enllà de l'estricta recopilació i dedica els seus esforços a proveir diferents avaluacions sobre la qualitat dels tests. Per la seva banda, una altra referència internacional important és Pro-Ed (<http://www.proedinc.com>), que edita també amb caràcter periòdic els seus *Test Publisher*, amb l'objectiu de recopilar els tests comercials disponibles, i *Test Critiques Publisher*, que funciona com un recurs complementari de l'anterior i també proporciona informació sobre la seva qualitat.

- **Bases de dades.** D'altra banda, en la Xarxa podem trobar alguns directoris oberts que proporcionen informació sobre els tests indexats, tant comercials com no comercials, i hi faciliten l'accés. Entre d'altres, podem destacar la *Test Collection* (http://www.ets.org/test_link) de l'Educational Testing Service, que inclou més de 25.000 tests. En aquest sentit, una altra font interessant és *Test Reviews Online* (<http://buros.unl.edu/buros/jsp/search.jsp>), una base de dades desenvolupada pel Buros Center for Testing que permet la cerca electrònica entre els tests incorporats a l'MMY i que ofereix accés comercial individual a les seves avaluacions.
- **Catàlegs de les editorials.** D'especial interès per a la cerca de tests en llengua no anglesa, i en especial en castellà, diferents editorials espanyoles publiquen els catàlegs de tests als quals donen llicència i proveeixen de la informació necessària per a conèixer-ne les característiques. Entre aquests, podem destacar els catàlegs de TEA Ediciones (<http://www.teaediciones.com>), del Grupo Albor-Cohs (<http://www.psicologia365.com>), de l'Instituto de Orientación Psicológica EOS (<http://www.eos.es>) i de Pearson (Pearson Clinical & Talent Assessment, <http://www.pearsonpsychcorp.es>).
- **Cerca de tests no comercials.** Finalment, hem de tenir present que no sempre els tests comercials disponibles en el mercat s'ajusten a les necessitats de l'avaluació. En aquest cas, en la cerca de tests no comercials per a utilitzar-los en l'avaluació psicològica, és convenient consultar les bases de dades que inclouen les publicacions científiques més rellevants en l'àmbit de coneixement de la psicologia. Més enllà de les recomanacions habituals, podem destacar *PsycTESTS* (<http://www.apa.org/pubs/databases/psyc-tests>), una base de dades especialitzada en tests no comercials desenvolupada per l'American Psychological Association i actualitzada amb una periodicitat mensual, en què es recopilen tests obtinguts dels mateixos autors, de publicacions en revistes científiques, llibres o tesis.

Una vegada trobat un o més tests potencialment interessants, és necessari avaluar-ne les característiques per a **valorar la conveniència del seu ús** amb relació als objectius de l'avaluació psicològica. Per a fer-ho, podem tornar al pro-

cés de construcció de tests exposat en l'apartat anterior i oferir alguns criteris a partir de les preguntes més importants que els professionals interessats en el seu ús han de formular-se amb relació als objectius del test, les seves propietats psicomètriques, el procés d'administració, el seu estatus i, finalment, l'adequació específica al procés d'avaluació:

- **Amb relació als objectius del test.** Entre d'altres, està clarament definit l'objectiu del test? Disposa del suport teòric necessari? Disposem de les evidències científiques necessàries per a utilitzar-lo com a instrument de mesura? S'actualitza regularment o, al menys, és objecte d'investigació per a valorar-ne les propietats psicomètriques i la utilitat d'acord amb els objectius que es proposa?
- **Amb relació a les seves propietats psicomètriques.** D'acord amb el procés de construcció, es proporciona la informació necessària sobre l'anàlisi dels ítems que componen el test? Quines evidències es proporcionen sobre la fiabilitat i la validesa de les mesures obtingudes mitjançant el test? Quins tipus de mostres, i amb quines grandàries, s'han utilitzat per a recollir aquestes evidències?
- **Amb relació al procés d'administració.** Disposa el test d'un manual adequat amb la informació i els materials necessaris per al seu ús? Té instruccions clares o plantilles per a la correcció? Si fos necessari, aquest manual disposa de normes o barems per a la interpretació de les puntuacions obtingudes?
- **Amb relació a l'estatus.** Es tracta d'un test comercial i, en cas de ser-ho, quin cost tenen el manual, les còpies del test, els fulls de resposta i la plantilla de correcció? És necessari algun tipus de qualificació o acreditació per a administrar-lo? Ha estat avaluat per alguna entitat o organisme independent?
- **Amb relació a l'adequació al procés d'avaluació.** Finalment, és pertinent d'acord amb els objectius d'avaluació? És el seu marc teòric congruent amb aquests objectius? Es dirigeix específicament a la població a la qual es vol administrar? Està adaptat o disposa d'una versió adaptada al context cultural en què vol ser utilitzat? S'ajusta a les condicions d'administració previstes en el procés d'avaluació, per exemple, tenint en compte el mitjà d'administració, els materials necessaris, el format de resposta als ítems o el temps requerit?

Avaluar un test abans d'utilitzar-lo és més que una pràctica convenient en l'exercici professional de la psicologia i, de fet, representa un repte important al qual les mateixes organitzacions professionals han dedicat diferents esforços per tal de sistematitzar-ne l'abordatge. Sense aturar-nos en les seves especificitats, i tenint en compte el diferent nivell de complexitat en què han estat desenvolupades, aquestes iniciatives es basen també en l'anàlisi de les de-

cisions preses en el procés de disseny i construcció dels tests, i proporcionen un marc d'avaluació de la qualitat dels tests disponibles en la literatura molt útil per a la valoració de la seva conveniència amb relació als objectius de l'avaluació psicològica.

Així, en el context internacional, hem de destacar la iniciativa de l'American Educational Research Association, l'American Psychological Association i el National Council on Measurement in Education (2014), que des de 1985 publiquen conjuntament els seus *The Standards for Educational and Psychological Testing*. D'altra banda, l'European Federation of Psychologists' Associations (2012) treballa en el desenvolupament de l'*EFPA Review model for the description and evaluation of psychological and educational tests* amb la intenció d'harmonitzar els criteris i processos d'avaluació de la qualitat dels tests a escala europea. Finalment, en el context estatal, hem de fer menció de la contribució del Consejo General de Colegios Oficiales de Psicólogos per a adaptar el model de l'EFPA a “**Un modelo para evaluar la calidad de los tests utilizados en España**” (Prieto i Muñiz, 2000). Aquest model ja ha estat utilitzat amb èxit per a avaluar la qualitat de deu dels tests més utilitzats pels psicòlegs espanyols (Muñiz, Fernández-Hermida, Fonseca-Pedrero, Campillo-Álvarez i Peña-Suárez, 2011).

3.3. Aspectes ètics i deontològics en l'ús de tests

Com hem discutit àmpliament al llarg d'aquesta introducció a la psicometria, l'ús de tests com a instruments d'avaluació psicològica és una de les pràctiques habituals en l'exercici de la psicologia. Tant al servei de la intervenció en els diferents camps professionals com en el cor del desenvolupament dels camps acadèmics i científics, els tests són uns instruments indispensables per a la mesura objectiva i estandarditzada dels fenòmens psicològics. I com a tals, es troben subjectes a una reflexió sobre les **conseqüències ètiques i morals que comporta el seu ús**. És a dir, en tant que instruments, poden ser utilitzats o no de manera adequada i d'acord amb unes finalitats o unes altres. No és aquest el lloc per a començar una discussió filosòfica en profunditat sobre el que significa actuar bé o fer les coses correctament, però sí per a concloure aquesta introducció reflexionant sobre algunes qüestions importants que tenen a veure amb un ús dels tests responsable, just i respectuós amb els drets i la dignitat de les persones.

Per fer-ho, repassarem alguns dels codis de conducta més importants desenvolupats en el camp de la psicologia que, en últim terme, presideixen el desenvolupament correcte de l'activitat professional, docent i investigadora. I analitzarem la manera en què aborden els aspectes ètics i morals vinculats a l'ús de tests en el context general de l'avaluació psicològica. Aquests codis són la resposta específica que la psicologia ha donat a un debat complex que, com d'altres disciplines, ha afrontat a partir de la Segona Guerra Mundial. Partint del Codi de Nuremberg (1947), la Declaració d'Hèlsinki (1964), l'Informe Belmont (1979) i les directrius proposades pel Council for International Organi-

⁽¹³⁾Per a una interessant discussió sobre els orígens podeu veure Israel i Hay (2006).

sations of Medical Sciences (1982), la consideració de les implicacions ètiques i morals del treball amb humans ha format i forma part de la pràctica quotidiana de professionals i investigadors¹³. En l'àmbit concret de la psicologia, són molts els codis que les organitzacions professionals dels diferents països han desenvolupat, però en aquest text ens referirem als treballs de l'American Psychological Association, l'European Federation of Psychologists' Associations i, en el cas espanyol, el Consejo General de Colegios Oficiales de Psicólogos. Finalment, comentarem unes directrius internacionals que, en referència al cas específic de l'ús dels tests, ha desenvolupat la International Test Commission.

Com una de les referències més importants a escala internacional, l'American Psychological Association ha contribuït al debat sobre les implicacions ètiques i morals en el camp de la psicologia declarant un dels primers codis deontològics. A partir dels anys cinquanta, i en revisions successives fins a l'actualitat, aquesta associació promou el que anomena ***Ethical principles of psychologists and code of conduct*** (American Psychological Association, 2010), en què estableix un preàmbul, cinc principis generals i deu estàndards que estableixen els límits i les pautes adequades per a una pràctica professional responsable. El preàmbul i els cinc principis es presenten com un posicionament ètic i moral, és a dir, un compromís amb els ideals de la psicologia i no tant com unes normes de compliment obligat. En aquest sentit, el codi reclama l'aspiració a la beneficència i no maleficència, la fidelitat i responsabilitat, la integritat, la justícia i el respecte pels drets i la dignitat de les persones.

En altres paraules, els psicòlegs han de fer tot el possible per a cercar i preservar el benestar de les persones, evitant produir cap mal i vetllant pels drets de les persones amb les quals interactuen en els diferents àmbits d'activitat professional. Així mateix, han de desenvolupar relacions de confiança, responsabilitzant-se del seu comportament i tractant de resoldre activament els conflictes d'interès que puguin presentar-se en l'exercici professional. A més a més, els psicòlegs han d'actuar honestament i amb rigor, evitant en tot moment l'engany o l'equívoc. I han de fer-ho també amb equitat i justícia, fent una atenció especial a assegurar que cap biaix ni límit en la seva competència o en la seva experiència professional contribueixi o doni suport a pràctiques injustes. Finalment, els psicòlegs han de vetllar pels drets de les persones a la privacitat, confidencialitat i dignitat, respectant les diferències culturals i individuals, i evitant contribuir o donar suport a qualsevol tipus de perjudici en les seves pràctiques professionals.

Per la seva banda, els estàndards es presenten com un conjunt de normes de compliment obligat per als membres de l'associació i els sotmet a les decisions o sancions del seu comitè ètic en cas de violació del codi en l'exercici professional. Aquests estàndards estan descrits de manera general i amb diferent nivell d'articulació, amb l'objectiu que puguin ser aplicats de manera transversal als diferents contextos d'activitat professional, docent o investigadora. Els estàndards remetent a deu aspectes fonamentals com són la resolució dels conflictes ètics, la competència professional, les relacions humanes, la privacitat i la

Per saber-ne més

L'estudiant interessat pot trobar a Campbell, Vasquez, Behne i Kinscherff (2010) un comentari detallat d'aquests estàndards i també uns dilemes a tall d'exemple de la seva aplicació en la pràctica.

confidencialitat, la publicitat i les contribucions públiques, el manteniment de la documentació professional i els honoraris, la formació, la investigació i la publicació dels resultats, l'avaluació i la teràpia.

En aquest sentit, i d'acord amb els objectius d'aquesta discussió sobre els aspectes ètics i deontològics vinculats a l'ús de tests, l'estàndard relatiu a l'avaluació descriu amb detall el marc general en què aquesta ha de ser portada a terme i planteja la necessitat de basar qualsevol judici o conclusió en les evidències obtingudes mitjançant l'ús apropiat dels instruments d'avaluació. A més a més, fa referència explícita a l'ús de tests i planteja la necessitat d'utilitzar tests vàlids i fiables, l'obligació de posar a disposició de les persones les seves respostes i la seva interpretació de manera confidencial, el desenvolupament de nous tests sempre d'acord amb els estàndards de qualitat de la psicometria, la desestimació de tests obsolets o no adequats per a la població objectiu i el compliment de les condicions i ajust a les llicències d'ús dels tests desenvolupats per tercers.

Per la seva banda, l'European Federation of Psychologists' Associations (2005) ha proposat un *Meta-code of ethics* que serveix com a marc de referència per al context europeu i cerca homogeneïtzar el tractament de les qüestions ètiques i deontològiques de les diferents organitzacions professionals dels estats membres. Per a fer-ho, aquest codi conté un preàmbul general en què estableix la seva obligació de vetllar pel desenvolupament professional dels psicòlegs europeus en totes aquestes qüestions i proposa quatre principis fonamentals que, articulats en un conjunt de recomanacions, han de regir als diferents codis, procediments i comitès ètics nacionals.

Així, el codi europeu planteja la defensa i el respecte dels drets i la dignitat de les persones, el reconeixement de la importància de la competència dels professionals que exerceixen la psicologia, la responsabilitat envers les persones, les comunitats i la societat en general i la promoció de la integritat en el desenvolupament de l'activitat professional, docent i investigadora. Tant els principis com les diferents especificacions es consideren interdependents, de manera que es planteja obertament la necessitat d'establir un debat i diàleg professional al davant de la complexitat dels conflictes ètics que es presenten en la pràctica. El contingut del seu articulat reclama l'atenció dels psicòlegs per qüestions importants relatives a l'exercici professional quotidià comparables a les establertes en el codi de l'American Psychological Association però, a diferència d'aquest, no fa cap referència explícita als processos d'avaluació psicològica ni a l'ús de tests en aquest context.

D'altra banda, amb relació a l'exercici professional en el camp de la psicologia a Espanya, el Consejo General de Colegios Oficiales de Psicólogos (2010) ha desenvolupat el seu *Codi deontològic* com la plasmació d'uns drets i deures professionals que, en últim terme, serveixen de base per al judici de la conducta dels seus col·legiats (Bermejo, 2009, per a una discussió de les modificacions més recents). Disposa d'un títol preliminar, a mode de preàmbul, que en

determina l'abast i emfatitza tant el respecte fonamental pel marc normatiu i jurídic com l'obligació de rebutjar qualsevol tipus d'impediment o limitació en l'exercici professional lliure i independent. També proposa uns principis generals que, a continuació, desenvolupa en un conjunt d'àrees fonamentals, i un annex final recull el reglament de la comissió deontològica estatal encarregada de vetllar per la correcta interpretació i aplicació del codi. Així, entre els principis generals que han de regir l'activitat professional dels psicòlegs a l'Estat espanyol planteja:

- 1) L'atenció primordial al benestar, la salut, la qualitat de vida i la plenitud del desenvolupament de les persones.
- 2) La protecció dels drets humans i la responsabilitat en la fonamentació objectiva i científica de les intervencions professionals.
- 3) El no-desenvolupament de pràctiques que atemptin contra la llibertat i la integritat física i psíquica de les persones o la contribució a aquestes pràctiques.
- 4) L'obligatorietat d'informar almenys els organismes col·legials si es coneix qualsevol violació dels drets humans o tractament degradant.
- 5) El respecte pels criteris morals i religiosos de les persones.
- 6) La no-discriminació per qualsevol diferència o motiu.
- 7) La denegació de qualsevol benefici o profit que es pugui treure com a conseqüència de les relacions de poder o superioritat establertes.
- 8) L'ús d'un llenguatge prudent i crític davant d'etiquetes menyspreadores o discriminatòries en els seus informes.
- 9) El respecte per l'activitat dels altres professionals i la lliure competència.
- 10) La denúncia de pràctiques il·legítimes o intrusives.
- 11) La imparcialitat davant dels possibles conflictes d'interès entre els psicòlegs o les institucions en què desenvolupen la seva activitat i els interessos de les persones.

L'articulat desenvolupa aquests principis en set àrees fonamentals com són la competència professional i la relació amb altres professionals, la intervenció, la investigació i la docència, l'obtenció i ús de la informació, la publicitat, els honoraris i la remuneració, i les garanties processals. Articulades amb diferents nivells de detall, aquestes àrees no preveuen específicament l'avaluació psicològica ni fan cap referència explícita a l'ús de tests. Aquestes qüestions queden relegades a un tractament genèric en les àrees de la competència professional i de l'obtenció i ús de la informació, de manera que es planteja la necessitat

de disposar dels coneixements necessaris per a l'ús dels mètodes, instruments, tècniques i procediments, sempre d'acord amb les evidències científiques necessàries que en garanteixin un ús adequat. Amb relació a l'obtenció i ús de la informació, circumscriu l'ús d'aquests mètodes, instruments, tècniques i procediments d'acord amb les condicions de confidencialitat i secret professional, i reconeix el dret de les persones a conèixer els seus resultats.

Finalment, més enllà dels codis deontològics professionals, hem de fer menció del treball específic que la International Test Commission (2000) ha desenvolupat en les seves *International guidelines for test use*. Atesa la disparitat en la normativa i el nivell de desenvolupament dels codis professionals dels diferents països, aquesta organització internacional planteja unes directrius específiques sobre l'ús dels tests i el seu paper en el context de l'avaluació psicològica que, entre d'altres, subscriu el Consejo General de Colegios Oficiales de Psicólogos. Tal com assenyalava en la introducció, la intenció d'aquestes directrius no és afegir un altre conjunt de recomanacions a les que ja hi ha, sinó que persegueix l'objectiu de crear una estructura coherent sota la qual es puguin entendre i aplicar els diferents codis i estàndards nacionals que desenvolupen els aspectes ètics i deontològics vinculats a l'ús de tests. Per fer-ho, estructura el seu articulat a partir d'un propòsit general que estableix un ús apropiat, professional i ètic dels tests, amb respecte pels drets de les persones, les raons per les quals s'utilitzen i el context en què s'apliquen.

Així mateix, com a gran eix conductor de les seves directrius, proposa la consecució d'aquest propòsit general mitjançant el desenvolupament i l'adquisició de les competències necessàries per a portar a terme l'administració de tests, la interpretació i comunicació adequades dels resultats, i la resolució de les dificultats, malentesos i conflictes que es puguin produir durant el procés. D'aquesta manera, aquestes directrius estableixen que els usuaris competents dels tests haurien de:

1) Responsabilitzar-se de l'ús ètic dels tests:

- Actuant de manera professional i ètica.
- Assegurant-se que tenen els coneixements i les habilitats necessàries.
- Fent-se responsables del seu ús dels tests.
- Mantenint de manera segura els materials dels tests.
- Garantint la confidencialitat dels resultats.

2) Comprometre's amb les bones pràctiques en l'ús dels tests:

- Estudiant la utilitat dels tests en els processos d'avaluació.
- Triant tests ben fonamentats i apropiats per a la situació.
- Atenent a les qüestions relacionades amb l'equitat.
- Preparant adequadament les condicions d'administració.
- Utilitzant correctament els tests.
- Obtenint les puntuacions i analitzant-les amb exactitud.

- Interpretant els resultats correctament.
- Comunicant els resultats de manera clara i exacta.
- Avaluant adequadament el funcionament i les propietats dels tests.

Totes aquestes directrius estan desenvolupades, al seu torn, en un conjunt de recomanacions concretes i representen un retrat complex de les dificultats i els reptes ètics i morals que planteja específicament l'ús dels tests com a instruments d'avaluació psicològica. Finalment, el text conclou amb uns annexos que desenvolupen algunes indicacions per al tractament específic de quatre qüestions importants com són el desenvolupament de polítiques i normatives sobre l'ús de tests en les organitzacions i empreses, la redacció de contractes entre les diferents parts involucrades en l'ús dels tests, l'administració de tests a persones amb discapacitats i la traducció de les pròpies directrius als diferents idiomes per part de les organitzacions estatals vinculades a l'exercici professional de la psicologia.

Com hem pogut fer notar, les qüestions ètiques i morals formen part intrínseca de l'exercici professional de la psicologia. Així ho reflecteixen els diferents codis deontològics que regeixen l'activitat dels psicòlegs als Estats Units, Europa i Espanya, mitjançant els quals es reconeix el repte que representa afrontar de manera activa, responsable i compromesa la complexitat de les situacions i contextos en què els professionals intervenen de manera quotidiana. Per aquesta raó és obligació dels psicòlegs exigir-se, i exigir als altres professionals amb qui col·laboren, un comportament exemplar davant dels dilemes ètics i morals que la pràctica pugui portar. Els tests, instruments imprescindibles per a l'avaluació i la intervenció psicològiques, no es troben al marge d'aquesta discussió malgrat que, com hem comentat, els codis no sempre fan un tractament específic sobre els processos d'avaluació i, concretament, sobre el paper que tenen els tests.

Com qualsevol altre tipus d'eina o instrument els tests poden ser emprats al servei d'unes finalitats o unes altres, però el compromís ha de ser la cerca constant i inequívoca del benefici i no perjudici de les persones, reaccionant davant de qualsevol tipus d'impediment que pugui limitar l'exercici professional honest, lliure i independent. En definitiva, respectant el dret de les persones a ser tractades i avaluades amb justícia, equitat i responsabilitat, i a conèixer els resultats i les conseqüències que de l'avaluació psicològica es puguin derivar. Com hem pogut veure, aquesta responsabilitat no es limita a l'administració de tests amb les garanties científiques necessàries sinó que també implica el domini de les teories, mètodes i tècniques que ha desenvolupat la psicometria per a comprendre'n el funcionament, valorar-ne la conveniència amb relació als objectius de l'avaluació i, finalment, evitar qualsevol tipus de perjudici que un ús inadequat pogués ocasionar. És per això que els psicòlegs han de ser curatsos amb la seva pròpia competència, la clau de volta per a un ús adequat dels

tests, amb l'objectiu de desenvolupar els coneixements i les habilitats necessàries que garanteixin un respecte profund pels drets i la dignitat de les persones a qui administren els seus tests.

Bibliografia

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, D. C.: American Psychological Association.

American Psychological Association (2010). Ethical principles of psychologists and code of conduct. Washington, D. C.: American Psychological Association. <<http://www.apa.org/ethics/code/principles.pdf>>

Anastasi, A. (1988). *Psychological testing* (6a. ed.). Nova York: Macmillan.

Bermejo, V. (2009). Motivaciones para la revisión y cambios en el código deontológico de la profesión de la psicología. *Papeles del Psicólogo*, 30 (3), 195-206. <<http://www.papelesdelpsicologo.es/pdf/1748.pdf>>

Bock, R. D. (1997). A brief history of Item response theory. *Educational Measurement: Issues and Practice*, 16 (4), 21-33.

Bock, R. D. i Jones, L. V. (1968). *The measurement and prediction of judgment and choice*. San Francisco: Holden-Day.

Bondy, M. (1974). Psychiatric antecedents of psychological testing (before Binet). *Journal of the History of the Behavioral Sciences*, 10 (2), 180-194.

Boring, E. G. (1978). *Historia de la psicología experimental*. Mèxic: Trillas.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in modern psychometrics*. Cambridge: Cambridge University Press.

Bowman, M. L. (1989). Testing individual differences in ancient China. *American Psychologist*, 44(3), 576-578.

Brennan, R. L. (1997). A perspective on the history of Generalizability Theory. *Educational Measurement: Issues and Practice*, 16 (4), 14-20.

Bridgman, P. W. (1927). *The logic of modern physics*. Nova York: Macmillan.

Buchanan, R. D. i Finch, S. J. (2005). History of psychometrics. A B. S. Everitt i D. Howell (Eds.). *Encyclopedia of statistics in behavioral science* (pp. 875-878). Chichester: Wiley.

Campbell, L., Vasquez, M., Behne, S., i Kinscherff, R. (2010). *APA Ethics code commentary and case illustrations*. Washington, DC: American Psychological Association.

Campbell, N. R. (1920). *Physics: The elements*. Cambridge: Cambridge University Press.

Campbell, N. R. (1928). *An account of the principles of measurement and calculation*. Londres: Longmans, Green, and co.

Carson, J. (1993). Army alpha, army brass, and the search for army intelligence. *Isis*, 84 (2), 278-309.

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin*, 40 (3), 153-193.

Chadha, N. K. (2009). *Applied psychometry*. Nova Delhi: Sage.

Consejo General de Colegios Oficiales de Psicólogos (2010). *Código deontológico*. Madrid: Consejo General de Colegios Oficiales de Psicólogos. <<http://www.cop.es/pdf/Codigo-Deontologico-Consejo-Adaptacion-Ley-Omnibus.pdf>>

Crocker, L. i Algina, J. (2006). *Introduction to classical and modern test theory*. Mason: CENGAGE Learning.

Cronbach, L. J., Rajaratnam, N., i Gleser, G. C. (1963). Theory of Generalizability: A liberation of reliability theory. *British Journal of Mathematical and Statistical Psychology*, 16 (2), 137-163.

Dingle, H. (1950). A theory of measurement. *The British Journal for the Philosophy of Science*, 1(1), 5-26.

Downing, S. M. (2006). Twelve steps for effective test development. A S. M. Downing i T. M. Haladyna (Eds.). *Handbook of test development* (pp. 3-25). Mahwah: Lawrence Erlbaum.

DuBois, P. H. (1970). *The history of psychological testing*. Boston: Allyn & Bacon.

European Federation of Psychologists' Associations (2005). *Meta-code of ethics*. Brussel·les: European Federation of Psychologists' Associations. <<http://www.efpa.eu/ethics/ethical-codes>>

European Federation of Psychologists' Associations (2012). *EFPA Review model for the description and evaluation of psychological and educational tests. Version 4.2.4*. Brussel·les: European Federation of Psychologists' Associations. <<http://www.efpa.eu/professional-development>>

Ferguson, A., Myers, C. S., Bartlett, R. J., Banister, H., Bartlett, F. C., Brown, W., Campbell, N. R., Craik, K. J. W., Drever, J., Guild, J., Houstoun, R. A., Irwin, J. O., Kaye, G. W. C., Philpott, S. J. F., Richardson, L. F., Shaxby, J. H., Smith, T., Thouless, R. H., i Tucker, W. S. (1940). Quantitative estimates of sensory events: Final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *Advancement of Science*, 1, 331-349.

Fernández-Ballesteros, R. (1997). Evaluación psicológica y tests. A A. Cordero (Ed.). *La evaluación psicológica en el año 2000* (pp. 11-26). Madrid: TEA Ediciones.

Fraser, C. O. (1980). Measurement in psychology. *British Journal of Psychology*, 71(1) 23-34.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *History of Psychology*, 87 (3), 564-567.

García Cueto, E. (1993). *Introducción a la psicometría*. Madrid: Siglo Veintiuno de España Editores.

Gibby, R. E., i Zickar, M. J. (2008). A history of the early days of personality testing in American industry: An obsession with adjustment. *History of Psychology*, 11(3), 164-184.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist*, 18 (8), 519-521.

Gleser, G. C., Cronbach, L. J., i Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, 30 (4), 395-418.

Goodenough, F. L. (1949). *Mental testing: Its history, principles, and applications*. Nova York: Rinehart.

Goslin, D. A. (1963). *The search for ability. Standardized testing in social perspective*. Nova York: Russell Sage Foundation.

Gruijter, D. N. M. de i Kamp, L. J. T. van der (2008). *Statistical test theory for the Behavioral Sciences*. Boca Raton: Chapman & Hall.

Guilford, J. P. (1936). *Psychometric methods*. Nova York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests*. Nova York: Wiley.

Gulliksen, H. (1968). Louis Leon Thurstone, experimental and mathematical psychologist. *History of Psychology*, 23(11), 786-802.

Hambleton, R. K. (1994). The rise and fall of criterion-referenced measurement? *Educational Measurement: Issues and Practice*, 13 (4), 21-26.

Hambleton, R. K., Swaminathan, H., i Rogers, H. J. (1991). *Fundamentals of Item response theory*. Newbury Park: Sage Publications.

Hand, D. J. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159 (3), 445-492.

Hilgard, E. R. (1987). *Psychology in America: A historical survey*. Orlando: Harcourt.

Holden, R. R. (2000). Psychometrics. A A. E. Kazdin (Ed.). *Encyclopedia of psychology* (vol. VI, pp. 417-419). Nova York: Oxford University Press.

International Test Commission (2000). *International guidelines for test use*. <<http://www.intestcom.org/upload/sitefiles/41.pdf>>

Kline, P. (1998). *The new psychometrics. Science, psychology, and measurement*. Londres: Routledge.

Jáñez, L. (1989). *Fundamentos de psicología matemática*. Madrid: Pirámide.

Jensen, A. R. (1980). *Bias in mental testing*. Nova York: Free Press.

Jones, L. V. (1971). The nature of measurement. A R. L. Thorndike (Ed.). *Educational measurement* (2a. ed.) (pp. 335-355). Washington, DC: American Council on Education.

Jones, L. V. i Thissen, D. (2007). A history and overview of Psychometrics. A C. R. Rao i S. Sinharay (Eds.). *Handbook of statistics 26. Psychometrics* (pp. 1-27). Amsterdam: Elsevier.

Kaplan, R. M. i Saccuzzo, D. P. (2001). *Psychological testing* (6a. ed.). Belmont: Wadsworth Publishing Company.

Lord, F. M. (1953). On the statistical treatment of football numbers. *The American Psychologist*, 8(12), 750-751.

Lord, F. M. i Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison Wesley.

Martin, O. (1997). La mesure en psychologie de Binet à Thurstone, 1900-1930. *Revue de synthèse*, 118(4), 457-493.

Matínez Arias, M. R., Hernández Lloreda, M. J., i Hernández Lloreda, M. V. (2006). *Psicometría*. Madrid: Alianza Editorial.

Maydeu-Olivares, A. i McArdle, J. J. (2005). *Contemporary psychometrics*. Mahwah: Lawrence Erlbaum.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *History of Psychology*, 100(3), 398-407.

Michell, J. (1999). *Measurement in psychology*. Cambridge: Cambridge University Press.

Muñiz, J. (1996). *Psicometría*. Madrid: Editorial Universitat.

Muñiz, J. (2003). *Teoría clásica de los tests*. Madrid: Pirámide.

Muñiz, J. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. *Papeles del Psicólogo*, 31 (1), 57-66. <<http://www.papelesdelpsicologo.es/pdf/1796.pdf>>

Muñiz, J., Fernández-Hermida, J. R., Fonseca-Pedrero, E., Campillo-Álvarez, A., i Peña-Suárez, E. (2011). Evaluación de tests editados en España. *Papeles del Psicólogo*, 32 (2), 113-128. <<http://www.papelesdelpsicologo.es/pdf/1947.pdf>>

Muñiz, J. i Fonseca-Pedrero, E. (2008). Construcción de instrumentos de medida para la evaluación universitaria. *Revista de Investigación en Educación*, 5, 13-25.

Muñiz, J. i Hambleton, R. K. (1992). Medio siglo de teoría de respuesta a los ítems. *Anuario de Psicología*, (52), 41-66.

Murphy, K. R. i Davidshofer, C. O. (2005). *Psychological testing* (6a. ed.). Upper Saddle River: Pearson education.

Nicolas, S. i Ferrand, L. (2002). Alfred Binet and higher education. *History of Psychology*, 5 (3), 264-283.

Padilla, M., Merino, J. M., Rodríguez-Miñón, P., i Moreno, E. (1996). *Psicología matemática I*. Madrid: UNED.

Popham, W. J. i Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement*, 6 (1), 1-9.

Prieto, G. i Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo*, 77, 65-75. <<http://www.papelesdelpsicologo.es/vernumero.asp?id=1102>>

Rogers, T. B. (1995). *The psychological testing enterprise*. Pacific Grove, CA: Brooks / Cole Publishing Company.

Rozeboom, W. W. (1966). Scaling theory and the nature of measurement. *Psychometrika*, 16(2), 170-233.

Rust, J. i Golombok, S. (2009). *Modern psychometrics. The science of psychological assessment* (3a. ed.). Londres: Routledge.

Samejima, F. (2000). Psychometric society. A. A. E. Kazdin (Ed.). *Encyclopedia of psychology* (vol. VI, pp. 419-420). Nova York: Oxford University Press.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 667-680.

Sokal, M. M. (1982). James McKeen Cattell and the failure of anthropometric mental testing, 1890-1901. A. W. R. Woodward i M. G. Ash (Eds.), *The problematic science. Psychology in nineteenth-century thought* (pp. 322-345). Nova York: Praeger Publishers.

Suppes, P. (1951). A set of independent axioms for extensive quantities. *Portugaliae Mathematica*, 10, 163-172.

Swistak, P. (1990). Paradigms of measurement. *Theory and Decision*, 29 (1), 1-17.

Thurstone, L. L. (1931). *The reliability and validity of tests: Derivation and interpretation of fundamental formulae concerned with reliability and validity of tests and illustrative problems*. Ann Arbor: Edwards Brothers.

Thurstone, L. L. (1947). *Multiple factor analysis: A development and expansion of the vectors of the mind*. Chicago: Chicago University Press.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, 16 (4), 8-14.

Torgerson, W. S. (1958). *Theory and methods of scaling*. Nova York: Wiley.

Urbina, S. (2004). *Essentials of psychological testing*. Hoboken: Wiley.

Valentine, E. R. (1999). The founding of the psychological laboratory, University College London: "Dear Galton...Yours truly, J. Sully." *History of Psychology*, 2 (3), 204-218.

Velleman, P. F. i Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *American Statistician*, 47 (1), 65-72.

Wolf, T. H. (1969). The emergence of Binet's conception and measurement of intelligence: A case history of the creative process. *Journal of the History of the Behavioral Sciences*, 5 (2), 113-134.

Yela, M. (1984). *Introducción a la teoría de los tests*. Madrid: Universidad Complutense.

Zeidner, J. i Drucker, A. J. (1988). *Behavioral science in the Army: A corporate history of the Army Research Institute*. Alexandria: United States Army Research Institute for the Behavioral and Social Sciences.

Fiabilitat

PID_00216188

Maite Barrios
Antoni Cosculluela

Temps mínim de dedicació recomanat: 4 hores



**Maite Barrios**

Professora agregada del Departament de Metodologia de les Ciències del Comportament de la Universitat de Barcelona (UB) i consultora dels Estudis de Psicologia i Ciències de l'Educació de la Universitat Oberta de Catalunya (UOC). Doctora en Psicologia per la UB. Desenvolupa la docència en l'àmbit de la metodologia impartint, entre d'altres, les assignatures de Psicomètria i Estadística aplicada a la psicologia i a informació i documentació. Coautora de diversos llibres i textos sobre estadística i tècniques de recerca. La seva recerca se centra, principalment, en la infometria, el biaix de gènere en ciència i els estudis bibliomètrics.

**Antoni Cosculluela**

Professor titular d'universitat del Departament de Metodologia de les Ciències del Comportament de la Facultat de Psicologia de la Universitat de Barcelona (UB) i consultor dels Estudis de Psicologia i Ciències de l'Educació de la Universitat Oberta de Catalunya (UOC). Doctor en Psicologia per la UB. La seva tasca docent s'ha desenvolupat en l'àmbit de la metodologia i els dissenys de recerca, i de l'anàlisi de dades en psicologia i en informació i documentació. Autor i coautor de diferents llibres d'estadística aplicada. Actualment la seva recerca s'orienta, entre altres camps, als estudis bibliomètrics en diferents àrees de coneixement.

Índex

Introducció.....	5
1. Concepte de fiabilitat segons la teoria clàssica.....	7
1.1. L'error de mesura	7
1.2. El coeficient de fiabilitat i la seva interpretació	8
1.3. Tipus d'errors de mesura	10
2. Equivalència de les mesures: mètode de les formes paral·leles.....	12
3. Estabilitat de les mesures: mètode test-retest.....	14
4. Consistència interna.....	15
4.1. Mètode de les dues meitats	15
4.1.1. Spearman-Brown	16
4.1.2. Rulon	17
4.1.3. Gutman-Flanagan	18
4.2. Covariació entre els ítems	19
4.2.1. Coeficient alfa de Cronbach	19
4.2.2. Inferències sobre α	22
4.2.3. Kuder-Richardson	26
5. Factors que afecten la fiabilitat.....	29
6. Estimació de la puntuació vertadera.....	33
6.1. Estimació de la puntuació vertadera a partir de la distribució normal de l'error aleatori	33
6.2. Estimació de la puntuació vertadera a partir del model de regressió lineal	34
7. Fiabilitat dels tests referits al criteri.....	36
7.1. Conceptes bàsics	36
7.2. Índexs d'acord que requereixen dues aplicacions del test	38
7.2.1. Coeficient de Hambleton i Novick	38
7.2.2. Coeficient kappa de Cohen	40
7.2.3. Coeficient de Livingston	41
7.3. Índexs d'acord que requereixen una única aplicació del test	42
7.3.1. Coeficient de Livingston (una única aplicació)	42
7.4. Fiabilitat interobservadors	43
7.4.1. Coeficient kappa	44
7.4.2. Coeficient de concordança	44

8. Estimació dels punts de tall.....	46
8.1. Mètodes basats en l'avaluació d'experts sobre els ítems	47
8.1.1. Mètode de Nedelsky	47
8.1.2. Mètode d'Angoff	49
8.1.3. Mètode del consens directe	50
8.2. Mètodes basats en l'avaluació d'experts sobre la competència dels subjectes	52
8.2.1. Mètode del grup de contrast	52
8.2.2. Mètode del grup límit	53
8.3. Mètodes de compromís	54
8.3.1. Mètode de Hofstee	54
8.3.2. Mètode de Beuk	56
Bibliografia.....	59

Introducció

En el llenguatge quotidià el terme *fiabilitat* s'associa a allò que funciona de manera correcta. Ens refiem del nostre despertador si sona a l'hora que s'ha programat, de la bàscula si ens proporciona sense error el nostre pes, fins i tot considerem que disposem d'un bon amic si sempre ens fa costat quan el necessitem. Si el despertador, la bàscula i el nostre amic no es comporten de la manera "correcta" considerem que no són fiables i en conseqüència decidim que no hi podem confiar.

En psicometria ens referim a la fiabilitat com la propietat que valora la consistència i precisió de la mesura. En conseqüència, si la mesura pren valors consistents i precisos, això ens permet confiar en els resultats obtinguts quan s'aplica un test. No obstant això, sabem que qualsevol procés de mesura (tant si es mesura un objecte físic com un aspecte psicològic) s'associa a algun grau d'error. La mesura perfecta no existeix. L'estudi de la fiabilitat d'un instrument de mesura ha de permetre saber fins a quin punt els resultats que s'obtenen a partir de la seva aplicació estan afectats per l'error que s'ha comès quan es mesura. Si l'error és petit, podem confiar en el resultats del test, si l'error és gran, el procés de mesurament deixa de tenir sentit. En aquest capítol s'aborda el tema de la fiabilitat des de dos vessants: d'una banda, es presenta la fiabilitat des de la perspectiva de la teoria clàssica dels tests (TCT) centrant-nos en els test referits a la norma (TRN), per a seguidament abordar el tema de la fiabilitat segons els tests referits al criteri (TRC).

Des de la perspectiva de la TCT es presenta el que s'entén per *error de mesura* i els diferents tipus d'error de mesura que es poden cometre en aplicar un test. Tot seguit, es descriu el model lineal proposat per Spearman i com a partir d'ell es deriva el coeficient de fiabilitat. Ens aturarem en com interpretar un coeficient de fiabilitat i també en les diferents estratègies que s'han anat desenvolupant per a calcular-lo: test-retest, formes paral·leles i consistència interna. A continuació, veiem tres dels factors que influeixen en la fiabilitat (variabilitat de les puntuacions obtingudes en el test, longitud del test i característiques dels ítems que el componen). Per acabar amb la TCT, es presenten dos procediments per a estimar la puntuació vertadera d'un subjecte: l'estimació que assumeix la distribució normal de l'error aleatori i l'estimació a partir del model de regressió lineal.

Veurem una manera diferent d'abordar la fiabilitat quan els tests que s'empren són instruments que tenen com a objectiu valorar la competència de les persones en algun domini concret del coneixement, els anomenats TRC. Per contextualitzar la fiabilitat en els TRC, en primer lloc, ens aturem en els conceptes bàsics que caracteritzen aquest tipus de tests i seguidament passem a descriure els tres procediments més clàssics per a abordar-ne la fiabilitat: els procedi-

ments que requireixen dues aplicacions del test per a valorar la consistència de la classificació, els que requireixen una única aplicació i aquells en què entra en joc el paper dels avaluadors. En el darrer apartat del mòdul es descriuen els mètodes que més freqüentment s'utilitzen per a determinar el punt de tall que permet una millor classificació entre els individus que són competents en el criteri d'interès i els que no ho són.

1. Concepte de fiabilitat segons la teoria clàssica

Segons la teoria clàssica dels tests, la fiabilitat d'un test està relacionada amb els errors de mesura aleatoris presents en les puntuacions obtingudes a partir de la seva aplicació. Així, un test serà més fiable com menys errors de mesura continguin les puntuacions obtingudes pels subjectes a qui s'aplica. Dit d'una altra manera, la fiabilitat d'un test serà la seva capacitat per a dur a terme mesures lliures d'errors.

1.1. L'error de mesura

Tot instrument de mesura ha de garantir, amb més o menys rigor, que les mesures que obtenim amb la seva aplicació es corresponen amb el vertader nivell o valor de la característica avaluada. Així, si volem mesurar la temperatura de l'aigua del mar un dia d'un calorós mes d'agost, necessitem un termòmetre que ens permeti obtenir aquesta dada. Si ho fem amb el termòmetre que vam comprar a uns grans magatzems per mesurar la temperatura de l'aigua de la banyera de casa, segurament obtindrem un valor que serà menys precís que si hi ho fem amb el termòmetre que utilitza el servei de meteorologia per a fer aquestes mesures. De totes maneres, segurament tant l'un com l'altre termòmetre mesuraran amb un cert grau d'imprecisió, possiblement més elevat en el primer cas que en el segon, però no exempt en cap del dos casos d'una certa desviació respecte a la vertadera temperatura de l'aigua. Si la mesura la féssim utilitzant un sofisticat instrumental cedit per la NASA, segurament ja tindríem força més garanties que la temperatura obtinguda es correspon amb molta més precisió amb la vertadera.

Per tant, qualsevol procés de mesura d'una característica dels objectes o dels subjectes té inherent un cert error en el seu mesurament. Podem trobar instruments de mesura amb més o menys capacitat per a minimitzar aquests errors, però difícilment podrem trobar-ne un que els elimini del tot.

En el nostre àmbit de la psicologia, en què les variables que mesurem habitualment són característiques pròpies dels subjectes, relacionades amb els seus trets de personalitat, les seves capacitats cognitives, els seus estats d'ànim, etc., i en què els instruments utilitzats per a aquest mesurament són generalment els tests, encara resulta més evident que les mesures que fem d'aquests atributs també estaran afectades per certs errors. Això farà que les puntuacions obtingudes amb les administracions d'aquests tests no es corresponguin exactament amb els vertaders nivells dels subjectes en la característica mesurada.

De totes maneres, alguns d'aquests errors propis de tot mesurament, poden respondre a factors sistemàtics que tindran una possible causa en el mateix procés de mesura, en l'instrument utilitzat o en certes característiques dels

objectes o subjectes mesurats. Així, si el termòmetre amb què mesuràvem la temperatura de l'aigua del mar té un error de construcció que fa que sempre mesuri un grau més del real, aquest error afectarà de la mateixa manera totes les mesures realitzades amb ell, i es podrà eliminar fent un bon calibratge de l'aparell. Altres errors no tenen aquest component sistemàtic, sinó que són aleatoris, indeterminats i no responen a cap factor que pugui ser conegut i, per tant, eliminat. Aquests errors aleatoris són els que estan implicats en el concepte de *fiabilitat*.

1.2. El coeficient de fiabilitat i la seva interpretació

Des de la teoria clàssica dels tests (TCT) de Spearman, es defineix el coeficient de fiabilitat d'un test $\rho_{xx'}$ com la correlació entre les puntuacions obtingudes per un grup de subjectes en dues formes paral·leles del test.

Segons la definició de *formes paral·leles d'un test* de la TCT, si un test tingués una fiabilitat perfecta, les puntuacions obtingudes per un subjecte en les dues formes paral·leles del test haurien de ser idèntiques i, per tant, la correlació entre les puntuacions d'un grup de subjectes en aquestes dues formes paral·leles del test seria 1 ($\rho_{xx'} = 1$). Qualsevol valor inferior a 1 haurà de ser degut als errors aleatoris propis de l'instrument de mesura.

A partir de la definició anterior del coeficient de fiabilitat, i tenint en compte els supòsits de la TCT, també podem expressar el coeficient de fiabilitat com el quocient entre la variància de les puntuacions veritables i la variància de les puntuacions empíriques. En aquest sentit, el coeficient de fiabilitat es pot interpretar com la proporció de variància de les puntuacions veritables (σ_v^2) que hi ha en la variància de les puntuacions empíriques (σ_x^2):

$$\rho_{xx'} = \frac{\sigma_v^2}{\sigma_x^2}$$

Podem deduir fàcilment, de l'expressió anterior i de les conseqüències derivades de la TCT, que aquest coeficient de fiabilitat serà igual a 1 menys la proporció de la variància dels errors (σ_e^2) que hi ha en la variància de les puntuacions empíriques.

$$\rho_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$$

Un índex directament relacionat amb el coeficient de fiabilitat és l'índex de fiabilitat (ρ_{xy}), que es defineix com la correlació entre les puntuacions empíriques d'un test i les puntuacions veritables. Aquest índex de fiabilitat és igual a l'arrel quadrada del coeficient de fiabilitat:

$$\rho_{xv} = \sqrt{\rho_{xx'}} = \frac{\sigma_v}{\sigma_x}$$

A l'hora d'interpretar el valor del coeficient de fiabilitat no hi ha un criteri únic i universalment acceptat com a adequat. Evidentment, valors propers a 0 denotaran una alta proporció de la variància dels errors en la variància de les puntuacions empíriques i, per tant, posaran de manifest que l'instrument utilitzat no és fiable, mentre que valors propers a 1 mostraran una baixa proporció de la variància dels errors en la variància de les puntuacions empíriques i, en conseqüència, ens permetran interpretar que el test utilitzat és fiable. Ara bé, el significat d'aquesta variància d'error difereix amb relació al tipus d'estratègia que s'ha utilitzat per a valorar la fiabilitat (aquestes estratègies es descriuen en els propers apartats). Cohen i Swerlik (2009) proposen que si s'ha utilitzat l'estratègia de test-retest la variància d'error serà deguda fonamentalment a les diferents administracions del test, si s'ha utilitzat l'estratègia de formes paral·leles l'error es pot atribuir a la construcció del test o bé a les diferents administracions i si s'ha valorat la fiabilitat a partir de la consistència del test la variància d'error pot ser deguda a la construcció del test.

Tret dels casos extrems, la determinació del valor mínim acceptable del coeficient de fiabilitat depèn de factors que poden influir en aquest valor, com la longitud del test o el procediment empíric o l'estratègia utilitzada per a calcular-lo, tal com s'ha comentat en el paràgraf anterior. De totes maneres, s'han intentat establir certs criteris generals que ens poden servir de referència. Així, en el seu text clàssic, Nunnally (1978) considera que el valor mínim acceptable del coeficient de fiabilitat estaria en 0,70, sobretot en un context de recerca bàsica. En canvi, en un context aplicat, com l'escolar o el clínic, la fiabilitat cal que sigui més elevada, situant-la per sobre de 0,80 o 0,90. En aquests àmbits és necessari tenir en compte que en aquests àmbits les conseqüències de la precisió dels instruments de mesura utilitzats poden ser decisives per als subjectes avaluats (pensem en els tests de diagnòstic clínic, o en els d'intel·ligència en població infantil per a determinar la necessitat de classes especials per als infants). Murphy i Davidshofer (2005) afirmen que en qualsevol context d'avaluació, una fiabilitat per sota de 0,6 es consideraria baixa i inacceptable. Kaplan i Saccuzo (2009) van una mica més enllà i suggereixen que coeficients de fiabilitat que oscil·len entre 0,7 i 0,8 són prou bons per a la majoria de les vegades en què els test s'utilitzen per a fins de recerca.

Altres autors consideren que un coeficient de fiabilitat molt proper a 1 pot significar que els ítems que componen el test són redundants en avaluar certs elements o factors del constructe mesurat i, per tant, no aporten informació rellevant respecte a altres elements o factors d'aquest constructe, la qual cosa tampoc no es pot considerar adequada.

Sense voler establir criteris estrictes i tenint en consideració tot el que s'ha exposat fins ara, podríem concloure que de manera general podem interpretar com una fiabilitat adequada valors del coeficient de fiabilitat dins l'interval de 0,70 a 0,95.

1.3. Tipus d'errors de mesura

Fins ara només ens hem referit a un tipus d'error: l'error de mesura, però cal mencionar que aquest no és l'únic error descrit en l'àmbit de la psicometria, sinó que també podem fer referència a l'error d'estimació, a l'error de substitució i a l'error de predicció.

Aquests errors estan relacionats amb les puntuacions dels subjectes considerades individualment. Així, l'error de mesura és, tal com el definirem tot seguit, la diferència entre la puntuació obtinguda per un subjecte en el test i la seva puntuació veritable en la característica mesurada per aquest test. Ara bé, si considerem els errors no individualment sinó amb relació a un grup o mostra de subjectes, podem obtenir els anomenats *errors típics*, que són les desviacions típiques d'aquests errors calculades a partir de les puntuacions de tots els subjectes de la mostra.

Podem doncs definir més formalment aquests diferents tipus d'errors, els seus errors típics associats i les fórmules que els expressen.

- **Error de mesura.** Definim l'error de mesura com la diferència entre la puntuació empírica d'un subjecte (X) i la seva puntuació veritable (V).

$$e = X - V$$

L'**error típic de mesura** és la desviació típica dels errors de mesura, i el podem expressar com:

$$\sigma_e = \sigma_x \sqrt{1 - \rho_{xx'}}$$

- **Error d'estimació de la puntuació veritable.** L'error d'estimació de la puntuació veritable es defineix com la diferència entre la puntuació veritable d'un subjecte i la seva puntuació veritable pronosticada mitjançant el model de la regressió (V').

$$e = V - V'$$

La desviació típica d'aquests errors d'estimació s'anomena **error típic d'estimació** i es pot obtenir amb l'expressió següent:

$$\sigma_{v,x} = \sigma_x \sqrt{1 - \rho_{xx'}} \sqrt{\rho_{xx'}} = \sigma_e \sqrt{\rho_{xx'}}$$

- **Error de substitució.** Es defineix l'error de substitució com la diferència entre les puntuacions d'un subjecte en dues formes paral·leles d'un test o, dit d'una altra manera, l'error que es comet en substituir la puntuació d'un subjecte en un test (X_1) per la puntuació obtinguda en una forma paral·lela d'aquest mateix test (X_2).

$$e = X_1 - X_2$$

S'anomena **error típic de substitució** la desviació típica dels errors de substitució, i el podem expressar de la manera següent:

$$\sigma_{e(s)} = \sigma_x \sqrt{1 - \rho_{xx'}} \sqrt{2}$$

- **Error de predicció.** L'error de predicció podem definir-lo com la diferència entre la puntuació d'un subjecte en un test (X_1) i la puntuació pronosticada en aquest test (X'_1) a partir d'una forma paral·lela X_2 . Seria l'error que cometríem si substituïssim la puntuació d'un subjecte en un test per la puntuació pronosticada a partir d'una forma paral·lela d'aquest test

$$e = X_1 - X'_1$$

En aquest sentit, X'_1 serà la puntuació pronosticada mitjançant la recta de regressió de X_1 sobre X_2 , i la podem expressar a partir del model lineal general adaptat a aquest context com:

$$X'_1 = \rho_{12} \frac{\sigma_1}{\sigma_2} (X_2 - \bar{X}_2) + \bar{X}_1$$

Definim l'**error típic de predicció** com la desviació típica dels errors de predicció, i el podem expressar com:

$$\sigma_{e(p)} = \sigma_x \sqrt{1 - \rho_{xx'}} \sqrt{1 + \rho_{xx'}}$$

2. Equivalència de les mesures: mètode de les formes paral·leles

De la definició i les fórmules del coeficient de fiabilitat no es pot extreure directament cap procediment que ens permeti calcular-ne el valor per a una determinada mostra de subjectes. Així, l'estimació empírica del valor del coeficient de fiabilitat cal obtenir-la mitjançant alguna estratègia que ens permeti o bé comparar les puntuacions dels mateixos subjectes en dues administracions del mateix test o en dues formes paral·leles del test, o bé analitzar les puntuacions d'un grup de subjectes en els diferents ítems del test.

Dels procediments empírics per a l'obtenció del coeficient de fiabilitat, el que es deriva directament de la TCT és l'anomenat **mètode de les formes paral·leles**, que consisteix en el càlcul del coeficient de correlació de Pearson entre les puntuacions d'una àmplia mostra de subjectes representativa de la població diana del test, en dues formes paral·leles d'un test prèviament obtingudes. Si, tal com es pot derivar de la definició de formes paral·leles d'un test, aquestes mesuren exactament el mateix constructe, exactament amb la mateixa precisió, les diferències que podem observar entre les puntuacions d'uns mateixos subjectes en les dues formes han de ser conseqüència dels errors de mesura del test i, per tant, aquest procediment ens proporcionarà un indicador adequat de la magnitud d'aquests errors de mesura, o sigui de la precisió o fiabilitat del test.

De fet, aquest indicador també serà representatiu del grau d'equivalència de les dues formes paral·leles del test, i en aquest sentit també rep el nom de **coeficient d'equivalència**.

La seva fórmula serà la del coeficient de correlació de Pearson aplicat a aquest cas:

$$r_{xx'} = r_{x_1x_2} = \frac{n \sum x_1x_2 - \sum x_1 \sum x_2}{\sqrt{[n \sum x_1^2 - (\sum x_1)^2][n \sum x_2^2 - (\sum x_2)^2]}}$$

Lectura de la fórmula

$r_{xx'}$: coeficient de fiabilitat del test.

$r_{x_1x_2}$: coeficient de correlació de Pearson

x_1 i x_2 : puntuacions obtingudes pels subjectes en cada una de les dues formes paral·leles del test.

Hem designat el coeficient de fiabilitat del test amb la notació de $r_{xx'}$ i no la de $\rho_{xx'}$ perquè estem obtenint aquest coeficient de fiabilitat per a una mostra concreta de subjectes, i per tant ens situem en un nivell empíric i no en un nivell teòric com hem fet en els apartats anteriors en presentar les bases i de-

⁽¹⁾ Aplicada a una mostra concreta de subjectes.

finicions de la fiabilitat i dels seus errors. Aquesta notació, que podem definir com a *mostral*¹, és la que continuarem utilitzant en els apartats següents, ja que en tots ells es presenten els procediments empírics per a l'obtenció del coeficient de fiabilitat i altres indicadors relacionats amb aquest.

La dificultat principal per a l'aplicació del mètode de les formes paral·leles recau en l'elaboració d'aquestes dues formes d'un test. Sovint resulta realment molt difícil construir dos tests que estiguin formats per ítems que puguin ser aparellats en funció de la seva total equivalència com requereix el concepte i la definició de *formes paral·leles*.

Els altres procediments per a l'obtenció del coeficient de fiabilitat es deriven dels dos vessants que podem considerar inherents al concepte de *fiabilitat*, l'estabilitat temporal i la consistència interna.

3. Estabilitat de les mesures: mètode test-retest

Un instrument de mesura fiable ha de proporcionar valors estables en diferents mesures dels mateixos subjectes obtingudes seqüencialment. Així, si mesurem un tret de personalitat suposadament estable d'uns mateixos subjectes en dos moments diferents amb el mateix test, el coeficient de correlació entre les seves puntuacions serà un bon indicador d'aquesta estabilitat del test i, per tant, de la seva fiabilitat.

Aquest mètode d'obtenció del coeficient de fiabilitat d'un test s'anomena **mètode test-retest** i consisteix en l'aplicació del test a una mateixa mostra de subjectes en dos moments diferents. Es calcula a partir del valor del coeficient de correlació de Pearson entre les puntuacions dels subjectes en aquests dos moments. La fórmula és exactament la mateixa que s'ha aplicat en l'apartat anterior per al cas de les formes paral·leles, únicament amb la diferència que en el test-retest x_1 i x_2 són les puntuacions obtingudes pels subjectes els dos cops que se'ls ha administrat el test.

L'avantatge d'aquest mètode és que no requereix dues formes diferents del test amb totes les dificultats que això comporta, i el seu inconvenient principal és que cal administrar dues vegades el mateix test als mateixos subjectes. Aquest fet pot comportar factors que distorsionin les puntuacions dels subjectes en la segona administració. Així, si els subjectes encara recorden el contingut del test, segurament el seu rendiment es veurà millorat respecte a la primera administració. En aquest sentit, un factor crucial per a una aplicació correcta d'aquest mètode és determinar l'interval temporal que cal deixar entre les dues administracions del test. Aquest interval temporal no pot ser ni tan curt que provoqui els efectes comentats anteriorment, ni tan ampli que es puguin donar canvis naturals (maduratiu, evolutiu o circumstancials) del tret o constructe mesurat que modifiquin les puntuacions dels subjectes en aquesta segona administració del test.

4. Consistència interna

Tal com s'ha dit anteriorment, un instrument de mesura fiable es caracteritza per una estabilitat temporal elevada i per una consistència interna adequada. La consistència interna fa referència al grau en què cada una de les parts de què es compon l'instrument és equivalent a la resta. Aquest principi aplicat al cas dels tests estarà determinat pel grau en què cada ítem, com a part bàsica constitutiva d'aquest, mostra una equivalència adequada a la resta d'ítems, o sigui que mesura amb el mateix grau el constructe mesurat. Així, si hi ha una equivalència elevada entre els ítems del test, se suposa que les respostes dels subjectes a aquests diferents ítems estaran altament correlacionades, i les diferents parts en què puguem dividir el test també mostraran aquesta elevada covariació.

Per posar un exemple, tot i les evidents diferències que es poden establir amb el cas dels tests, la consistència interna d'una cinta mètrica queda garantida si cada una de les seves parts (suposem els diferents centímetres que la componen) és equivalent a la resta. Així, podrem dividir la cinta en diferents parts iguals (dues, tres etc.), i cada una mesurarà exactament el mateix. Evidentment, aquesta exactitud en la consistència interna de la cinta mètrica és pràcticament impossible d'assolir en el cas dels tests, però l'exemple pot servir per a situar adequadament el concepte de *consistència interna* referit al cas de la construcció d'instruments de mesura en psicologia.

4.1. Mètode de les dues meitats

Pot derivar-se fàcilment a partir del que s'ha exposat en l'apartat anterior que, si dividim un test en dues meitats, aquestes han de ser equivalents per a garantir una consistència interna adequada. El grau d'equivalència de les dues meitats es pot avaluar calculant la correlació entre les puntuacions dels subjectes en aquestes dues meitats. Així, la correlació entre les puntuacions d'un grup de subjectes en les dues meitats en què podem dividir un test serà un indicador del grau de la seva consistència interna, i per tant de la seva fiabilitat. Aquest és el principi en què es basa el **mètode de les dues meitats** que presenta l'avantatge respecte als mètodes anteriors que requereix una sola aplicació del test en una mostra de subjectes.

A l'hora de decidir com fer aquesta partició del test en dues meitats, cal tenir en compte que si ho fem, per exemple, deixant els primers ítems en una meitat i els darrers en l'altra, poden posar-se en joc factors que alterin l'equivalència del rendiment dels subjectes en les dues meitats. Així, és conegut que els subjectes acostumen a prestar més atenció als primers ítems d'un test, amb la millora consegüent del seu rendiment, o són més sincers en les respostes. Aquests possibles factors incidirien en una manca de consistència interna del test, no producte dels errors de mesura aleatoris propis de la fiabilitat del test, sinó

d'errors de mesura sistemàtics independents d'aquesta fiabilitat, ja que una de les dues meitats del test estaria afavorida per un rendiment millor o més acurat dels subjectes.

Per a evitar factors com els comentats anteriorment, habitualment es divideix el test en dues meitats, deixant els ítems parells en una meitat i els imparells en l'altra. Amb aquest procediment s'eviten bona part d'aquests factors i es garanteix de manera més probable l'equivalència entre les dues meitats.

4.1.1. Spearman-Brown

Com es comentarà més abastament en els propers apartats, el nombre d'ítems que componen un test incideix en la seva fiabilitat. Així, essent constants altres factors, com més ítems conté un test més elevada n'és la fiabilitat. Aquest efecte de la longitud d'un test sobre el coeficient de fiabilitat cal tenir-lo present en aplicar el mètode de les dues meitats. Per tant, si calculem el coeficient de correlació entre el total de les puntuacions dels subjectes en els ítems parells d'una banda i, d'una altra, el total de les seves puntuacions en els ítems imparells, i a partir d'aquest coeficient de correlació quantifiquem la fiabilitat del test, aquest valor estarà esbiaixat negativament, ja que el calculem a partir de la correlació entre la meitat del nombre total d'ítems del test. Aquest fet significa que cal corregir aquest coeficient de correlació per a obtenir el coeficient de fiabilitat de la totalitat del test. Aquesta correcció és denominada *de Spearman-Brown*, i és un cas concret de la fórmula del mateix nom que s'aplica per a obtenir la fiabilitat d'un test un cop aquest s'ha allargat o escurçat, afegint o eliminant una determinada quantitat d'ítems:

La fórmula per a l'obtenció del coeficient de fiabilitat d'un test a partir del mètode de les dues meitats amb la correcció de Spearman-Brown és:

$$r_{xx'} = \frac{2r_{pi}}{1+r_{pi}}$$

Exemple

En la taula següent tenim les puntuacions de vuit subjectes en un test de sis ítems dicotòmics:

Taula 1

Subjectes	Ítems					
	1	2	3	4	5	6
A	1	1	1	1	0	1
B	0	1	1	1	1	0
C	1	1	0	1	1	0
D	1	1	1	1	1	1
E	1	1	1	1	1	1

Lectura de la fórmula

$r_{xx'}$: coeficient de fiabilitat del test.

r_{pi} : coeficient de correlació de Pearson entre el sumatori de les puntuacions dels ítems parells i les dels ítems imparells.

Subjectes	Ítems					
	1	2	3	4	5	6
F	0	1	1	0	0	0
G	0	1	1	0	1	0
H	1	0	1	0	0	0

Per obtenir el coeficient de fiabilitat del test pel mètode de les dues meitats, en primer lloc calculem el sumatori de les puntuacions en els ítems parells per cada subjecte d'una banda, i de l'altra el sumatori dels seus ítems imparells, i obtenim el coeficient de correlació entre aquestes dues distribucions de valors:

Taula 2

Subjectes	Ítems parells	Ítems imparells
A	3	2
B	2	2
C	2	2
D	3	3
E	3	3
F	1	1
G	1	2
H	0	2

Aplicant la fórmula del coeficient de correlació de Pearson entre els ítems parells i els imparells, obtenim un coeficient igual a 0,62 ($r_{pi} = 0,62$).

Un cop calculat aquest valor, apliquem la fórmula de Spearman-Brown per obtenir el coeficient de fiabilitat del test:

$$r_{xx'} = \frac{2r_{pi}}{1+r_{pi}} = \frac{2 \times 0,62}{1+0,62} = 0,76$$

El coeficient de fiabilitat del test és de 0,76.

4.1.2. Rulon

La fórmula de Rulon (1939) per a calcular la fiabilitat d'un test també parteix de la seva divisió en dues meitats. Es basa en el supòsit que, si les dues meitats són paral·leles, les puntuacions dels subjectes en cada una només poden diferir com a conseqüència dels errors aleatoris. Per tant, la variància de les diferències entre aquestes dues meitats serà una estimació de la variància dels errors, i podrem substituir la variància dels errors de la fórmula del coeficient de fiabilitat derivada de la TCT per la variància de les diferències.

Així, doncs, la fórmula de Rulon és la següent:

$$r_{xx'} = 1 - \frac{S_d^2}{S_x^2}$$

Exemple

Si apliquem la fórmula de Rulon a l'exemple de l'apartat anterior necessitarem calcular:

Taula 3

Subjectes	Ítems parells	Ítems imparells	Diferència P-I	Total
A	3	2	1	5
B	2	2	0	4
C	2	2	0	4
D	3	3	0	6
E	3	3	0	6
F	1	1	0	2
G	1	2	-1	3
H	0	2	-2	2

Després de calcular la variància de les diferències ($S_d^2 = 0,6875$), i la variància de les puntuacions totals ($S_x^2 = 2,25$), podem aplicar la fórmula de Rulon:

$$r_{xx'} = 1 - \frac{S_d^2}{S_x^2} = 1 - \frac{0,6875}{2,25} = 0,69$$

El coeficient de fiabilitat del test, obtingut amb la fórmula de Rulon és de 0,69.

4.1.3. Gutman-Flanagan

Tant Flanagan (1937) com Gutman (1945) van obtenir una fórmula equivalent a la de Rulon, a partir de les variàncies dels ítems parells i imparells. Es basa en el mateix principi que l'anterior, però resulta més senzilla d'obtenir:

$$r_{xx'} = 2 \left(1 - \frac{S_p^2 + S_i^2}{S_x^2} \right)$$

Exemple

Aplicant la fórmula de Gutman-Flanagan a l'exemple anterior, tenim:

Taula 4

Subjectes	Ítems parells	Ítems imparells	Total
A	3	2	5
B	2	2	4
C	2	2	4

Lectura de la fórmula

S_d^2 : variància de les diferències entre les puntuacions dels subjectes en les dues meitats del test.

S_x^2 : variància de les puntuacions totals dels subjectes en el test.

Lectura de la fórmula

S_p^2 : variància de les puntuacions dels subjectes en els ítems parells del test.

S_i^2 : variància de les puntuacions dels subjectes en els ítems imparells del test.

S_x^2 : variància de les puntuacions totals dels subjectes en el test.

Subjectes	Ítems parells	Ítems imparells	Total
D	3	3	6
E	3	3	6
F	1	1	2
G	1	2	3
H	0	2	2

Calculant les diferents variàncies obtenim:

Variància de les puntuacions en els ítems parells: $S_p^2 = 1,11$.

Variància de les puntuacions en els ítems imparells: $S_i^2 = 0,36$.

Variància de les puntuacions totals en el test: $S_x^2 = 2,25$.

I per tant:

$$r_{xx'} = 2 \left(1 - \frac{S_p^2 + S_i^2}{S_x^2} \right) = 2 \left(1 - \frac{1,11 + 0,36}{2,25} \right) = 0,69$$

Com podem observar, el valor del coeficient de fiabilitat, calculat a partir de l'expressió de Gutman-Flanagan, és exactament igual a l'obtingut amb la fórmula de Rulon, com no podia ser d'altra manera, ja que com hem dit les dues fórmules són equivalents. Tant l'una com l'altra proporcionen un coeficient de fiabilitat del test igual a 0,69.

4.2. Covariació entre els ítems

Si, tal com s'ha comentat anteriorment, la consistència interna d'un test fa referència al grau en què cada una de les parts o ítems de què es compon és equivalent a la resta, la covariació entre aquests ítems també ens proporcionarà un indicador adequat d'aquesta consistència interna. De fet, aquest procediment és una extensió dels procediments anteriors de la divisió del test en dues meitats, en el cas límit de dividir-lo en tantes parts com ítems el componen. Així, cada ítem representarà una part equivalent del conjunt de tots, o sigui del test o escala total. De la mateixa manera que les dues parts del test han de mantenir una correlació elevada entre elles per a garantir la consistència interna del conjunt, cada ítem també ha de mostrar una covariació adequada a la resta dels ítems per a assolir la mateixa garantia de consistència interna.

4.2.1. Coeficient alfa de Cronbach

El coeficient alfa (α) de Cronbach (1951) expressa la consistència interna d'un test, a partir de la covariació entre els seus ítems. Com més elevada sigui la proporció de la covariació entre aquests ítems respecte a la variància total del test, més elevat serà el valor del coeficient alfa (α) de Cronbach, i més elevada la seva fiabilitat.

Hi ha diferents fórmules per a obtenir el valor del coeficient alfa, essent la més àmpliament utilitzada la que es deriva del càlcul de les variàncies de cada ítem i de la variància de les puntuacions totals en el test.

Aquesta fórmula és la següent:

$$\alpha = \frac{n}{n-1} \left[1 - \frac{\sum_{j=1}^n S_j^2}{S_x^2} \right]$$

En el nostre exemple:

Taula 5

Subjectes	Ítems						x
	1	2	3	4	5	6	
A	1	1	1	1	0	1	5
B	0	1	1	1	1	0	4
C	1	1	0	1	1	0	4
D	1	1	1	1	1	1	6
E	1	1	1	1	1	1	6
F	0	1	1	0	0	0	2
G	0	1	1	0	1	0	3
H	1	0	1	0	0	0	2
Variàncies	0,234	0,109	0,109	0,234	0,234	0,234	2,25

$$\alpha = \frac{n}{n-1} \left[1 - \frac{\sum_{j=1}^n S_j^2}{S_x^2} \right] = \frac{6}{5} \left(1 - \frac{0,234 + 0,109 + 0,109 + 0,234 + 0,234 + 0,234}{2,25} \right) = 0,583$$

D'altra banda, la fórmula del coeficient alfa que es deriva directament de la covariància entre els diferents ítems és expressada per:

$$\alpha = \frac{n}{n-1} \left[\frac{\sum_{j \neq k}^n \text{cov}(j, k)}{S_x^2} \right]$$

Així, en l'exemple exposat anteriorment, calcularem la variància de les puntuacions totals en el test (x), i les covariàncies entre els diferents ítems:

Lectura de la fórmula

n : nombre d'ítems del test.

$\sum_{j=1}^n S_j^2$: sumatori de les variàncies dels n ítems.

S_x^2 : variància de les puntuacions totals en el test.

Lectura de la fórmula

n : nombre d'ítems del test.

$\sum_{j \neq k} \text{cov}(j, k)$: sumatori de les covariàncies dels n ítems.

S_x^2 : variància de les puntuacions en el test.

Taula 6

Subjectes	Ítems						x
	1	2	3	4	5	6	
A	1	1	1	1	0	1	5
B	0	1	1	1	1	0	4
C	1	1	0	1	1	0	4
D	1	1	1	1	1	1	6
E	1	1	1	1	1	1	6
F	0	1	1	0	0	0	2
G	0	1	1	0	1	0	3
H	1	0	1	0	0	0	2

Les covariàncies entre els sis ítems són:

Taula 7

	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6
Ítem 1		-0,047	-0,047	0,109	-0,016	0,141
Ítem 2	-0,047		-0,016	0,078	0,078	0,047
Ítem 3	-0,047	-0,016		-0,047	-0,047	0,047
Ítem 4	0,109	0,078	-0,047		0,109	0,141
Ítem 5	-0,016	0,078	-0,047	0,109		0,016
Ítem 6	0,141	0,047	0,047	0,141	0,016	

i el seu sumatori:

$$\sum \sum \text{cov}(j, k) = -0,047 + -0,047 + 0,109 + \dots + 0,047 + 0,141 + 0,016 = 1,094$$

mentre que la variància de les puntuacions totals: $S_x^2 = 2,25$.

Per tant:

$$\alpha = \frac{n}{n-1} \left[\frac{\sum \sum_{j \neq k}^n \text{cov}(j, k)}{S_x^2} \right] = \frac{6}{5} \left(\frac{1,094}{2,25} \right) = 0,583$$

La fórmula del coeficient alfa també es pot expressar en funció del quocient entre la mitjana de les covariàncies i la mitjana de les variàncies dels diferents ítems del test. Aquest quocient, que designen com a r_1 , constitueix una estimació de la fiabilitat de cada ítem. En aquest sentit, la fórmula del coeficient alfa a partir de r_1 és una aplicació de la correcció de Spearman-Brown, que hem

comentat per al cas de les dues meitats, a partir de l'estimació de la fiabilitat de cada ítem, tenint en compte que si tenim n ítems és com si haguéssim allargat n cops l'ítem inicial.

$$\alpha = \frac{n(r_1)}{1 + (n-1)r_1}$$

Per al nostre exemple, la mitjana de les covariàncies és d'1,094/30, mentre que la mitjana de les variàncies és d'1,154/6.

Per tant, $r_1 = \frac{1,094/30}{1,154/6} = 0,189$ i el valor de α :

$$\alpha = \frac{n(r_1)}{1 + (n-1)r_1} = \frac{6 \times 0,189}{1 + (5 \times 0,189)} = 0,583$$

Com podem observar i com no podia ser d'altra manera, totes les diferents fórmules de càlcul del coeficient alfa de Cronbach aplicades a les dades del nostre exemple ens proporcionen el mateix valor.

4.2.2. Inferències sobre α

Un cop hem obtingut el valor del coeficient alfa de Cronbach per a una mostra determinada de subjectes, podem estar interessats a comprovar-ne la significació estadística o a determinar entre quins valors pot fluctuar aquest coeficient en la població. D'altra banda, també pot interessar-nos comparar dos coeficients alfa obtinguts en dues mostres independents o bé obtinguts en la mateixa mostra per a determinar si la diferència entre ells és estadísticament significativa.

Contrast per un sol coeficient

Kristof (1963) i Feldt (1965) van proposar un estadístic de contrast per tal de comprovar si un determinat valor del coeficient alfa pot ser compatible amb un cert valor poblacional. Així, podem analitzar si aquest valor d'alfa és estadísticament significatiu, o sigui si podem descartar la hipòtesi segons la qual el seu valor en la població és 0, o bé podem analitzar si aquest valor difereix significativament o no d'un determinat valor fixat prèviament en la població.

L'estadístic de contrast és:

$$F = \frac{1 - \alpha}{1 - \hat{\alpha}}$$

I es distribueix segons una distribució F de Snedecor amb $(N - 1)$ i $(n - 1)(N - 1)$ graus de llibertat.

Podem aplicar aquest estadístic de contrast al nostre exemple del test de sis ítems que ens ha donat un valor d'alfa de 0,583, obtingut en una mostra de vuit subjectes.

La hipòtesi nul·la que plantejarem és la que aquest coeficient alfa és igual a 0 en la població, cas més habitual i que comporta la no significació estadística d'aquest coeficient, mentre que l'alternativa representarà la desigualtat respecte al valor 0, i per tant la seva significació estadística.

Els passos a seguir en aquest contrast seran:

Hipòtesi nul·la: $\alpha = 0$

Hipòtesi alternativa: $\alpha \neq 0$

Càlcul de l'estadístic de contrast:

$$F = \frac{1 - \alpha}{1 - \hat{\alpha}} = \frac{1 - 0}{1 - 0,583} = 2,398$$

Valors crítics de la distribució F de Snedecor amb 7 $(N - 1)$ i 35 $((n - 1)(N - 1))$ graus de llibertat, per a un nivell de confiança del 95% i contrast bilateral són:

$$F_{0,975(7,35)} \approx 2,62 \text{ i } F_{0,025(7,35)} \approx 0,23^2$$

Com que el valor de l'estadístic de contrast obtingut (2,398) es troba dins de l'interval comprès entre els valors crítics 2,62 i 0,23, acceptem la hipòtesis nul·la i podem concloure que, a partir de les nostres dades i amb un nivell de confiança del 95%, no tenim prou evidències per a descartar que el valor del coeficient alfa en la població és de 0, i per tant aquest coeficient no és estadísticament significatiu.

Com a derivació senzilla del que s'ha exposat en aquest apartat, també podem determinar l'interval de confiança pel valor del coeficient alfa obtingut. En aquest sentit, en la fórmula de l'estadístic de contrast, només cal substituir els valors crítics de la distribució F i aïllar els valors de α :

$$\frac{1 - \alpha}{1 - 0,583} \leq 2,62$$

$$\alpha \geq 1 - 2,62(1 - 0,583) = -0,09$$

$$\frac{1 - \alpha}{1 - 0,583} \geq 0,23$$

$$\alpha \leq 1 - 0,23(1 - 0,583) = 0,90$$

⁽²⁾Nota

$$F_{0,025(7,35)} = \frac{1}{F_{0,975(35,7)}} \approx \frac{1}{4,31} = 0,23$$

La interpretació d'aquests valors anirà en el sentit de considerar que, amb un nivell de confiança del 95%, els valors del coeficient alfa en la població estaran compresos entre -0,09 i 0,90. Un cop establert aquest interval de confiança, podríem resoldre l'acceptació o no de qualsevol valor del coeficient en la hipòtesi nul·la del contrast corresponent. Així, si el valor del coeficient alfa poblacional plantejat en la hipòtesi nul·la cau dins de l'interval de confiança,

Lectura de la fórmula

N : nombre de subjectes.

n : nombre d'ítems.

α : valor d'alfa en la població.

$\hat{\alpha}$: valor d'alfa calculat en la mostra.

no podem descartar la certesa d'aquesta hipòtesi nul·la, mentre que si no hi cau, podem descartar-la i acceptar la hipòtesi alternativa per al nivell de confiança establert.

En el nostre exemple, com que el valor de 0 del coeficient està comprès en l'interval (entre -0,09 i 0,90) no podem rebutjar la hipòtesi nul·la, tal com ja hem vist anteriorment. Es deriva directament de l'anterior el fet que si un coeficient alfa empíricament obtingut no es troba comprès en l'interval de confiança construït, queda determinada la seva significació estadística sense necessitat de dur a terme el contrast per una hipòtesi nul·la igual a 0.

Contrast per a dos coeficients en mostres independents

També podem estar interessats a comprovar si dos coeficients alfa obtinguts en mostres diferents de subjectes són iguals o no. Per respondre aquesta qüestió aplicarem un contrast per a dos coeficients en mostres independents. Feldt (1969) va proposar l'estadístic w , que permet determinar si la diferència entre els dos coeficients és estadísticament significativa:

$$w = \frac{1 - \hat{\alpha}_1}{1 - \hat{\alpha}_2}$$

en què w es distribueix segons una F de Snedecor amb $(N_1 - 1)$ i $(N_2 - 1)$ graus de llibertat.

Podem dur a terme aquest contrast per al nostre exemple amb un nivell de confiança del 95%, suposant que s'ha administrat el mateix test a una mostra de deu subjectes i que hem obtingut un coeficient alfa de 0,65. Farem el contrast seguint aquests passos:

Hipòtesi nul·la: $\hat{\alpha}_1 = \hat{\alpha}_2$

Hipòtesi alternativa: $\hat{\alpha}_1 \neq \hat{\alpha}_2$

Càlcul de l'estadístic de contrast: $w = \frac{1 - \hat{\alpha}_1}{1 - \hat{\alpha}_2} = \frac{1 - 0,583}{1 - 0,65} = 1,19$

Valors crítics de la distribució F de Snedecor amb 7 (N_1) i 9 (N_2) graus de llibertat, per a un nivell de confiança del 95% i contrast bilateral són:

$$F_{0,975(7,9)} = 4,20 \text{ i } F_{0,025(7,9)} = 0,21$$

Lectura de la fórmula

$\hat{\alpha}_1$ i $\hat{\alpha}_2$: coeficients obtinguts en cada una de les dues mostres.

N_1 i N_2 : grandàries d'aquestes mostres.

Com que el valor de l'estadístic de contrast obtingut (1,19) cau dins de l'interval comprès entre els valors crítics (0,21 i 4,20), no tenim prou evidències per a rebutjar la hipòtesi nul·la, i per tant hem de concloure que la diferència entre els dos coeficients no és estadísticament significativa.

Contrast per a dos coeficients en mostres dependents

És habitual que els dos coeficients alfa obtinguts s'hagin calculat a partir de la mateixa mostra de subjectes. Més poc freqüent és que s'hagin obtingut a partir de dues mostres de subjectes relacionats entre ells per algun criteri d'aparellament (per exemple, parelles de bessons, pare-mare). Tanmateix, tant en un supòsit com en l'altre, anomenen el disseny *demostrés dependents*. Seria el cas, per exemple, si apliquéssim un disseny experimental de mesures repetides, i administréssim el mateix test a un sol grup de subjectes en dos moments diferents. En aquest sentit, podríem comparar els dos coeficients alfa obtinguts per a determinar la possible diferència estadísticament significativa entre ells.

Feldt (1980) va proposar un estadístic de contrast per a comparar dos coeficients alfa obtinguts en mostres dependents:

$$t = \frac{(\hat{\alpha}_1 - \hat{\alpha}_2)\sqrt{N-2}}{\sqrt{4(1-\hat{\alpha}_1)(1-\hat{\alpha}_2)(1-r_{12})}}$$

Per a il·lustrar l'aplicació d'aquest contrast podem considerar que el test utilitzat en els apartats anteriors l'hem administrat de nou una segona vegada a la mateixa mostra de vuit subjectes.

En la taula següent tenim les puntuacions d'aquests subjectes les dues vegades en què se'ls ha administrat el test.

Taula 8

	Ocasíó 1							Ocasíó 2						
	Ítems							Ítems						
Subjectes	1	2	3	4	5	6	x_1	1	2	3	4	5	6	x_2
A	1	1	1	1	0	1	5	1	1	0	1	0	1	4
B	0	1	1	1	1	0	4	0	1	0	1	1	0	3
C	1	1	0	1	1	0	4	1	1	0	1	1	0	4
D	1	1	1	1	1	1	6	1	1	0	1	1	1	5
E	1	1	1	1	1	1	6	1	1	0	1	1	1	5
F	0	1	1	0	0	0	2	0	1	0	0	0	0	1
G	0	1	1	0	1	0	3	0	1	0	0	1	0	2
H	1	0	1	0	0	0	2	0	1	0	0	0	0	1

Lectures recomanades

Woodruff i Feldt (1986) van fer extensiu el contrast anterior al cas de més de dos coeficients comparats simultàniament. No ens estem en aquesta qüestió, ja que la seva aplicació no és tan freqüent, però remetem l'estudiant interessat als textos de Muñiz (2003), i Barbero, Vila i Suárez (2003), en què pot trobar el seu desenvolupament clarament explicat.

Lectura de la fórmula

t : es distribueix segons una distribució t de Student amb $N - 2$ graus de llibertat.

$\hat{\alpha}_1$ i $\hat{\alpha}_2$: valors dels dos coeficients alfa.

N : nombre de subjectes de la mostra.

r_{12} : correlació entre les puntuacions dels subjectes en les dues administracions del test.

A partir d'aquestes dades, calculem el coeficient alfa per a la segona administració del test ($\hat{\alpha}_2 = 0,718$), i la correlació entre les puntuacions totals dels subjectes aquestes dues vegades ($r_{12} = 0,98$).

El contrast, per a determinar la possible diferència estadísticament significativa entre els dos coeficients alfa, amb un nivell de confiança del 95%, seguirà els passos següents:

Hipòtesi nul·la: $\hat{\alpha}_1 = \hat{\alpha}_2$

Hipòtesi alternativa: $\hat{\alpha}_1 \neq \hat{\alpha}_2$

Càlcul de l'estadístic de contrast:

$$t = \frac{(\hat{\alpha}_1 - \hat{\alpha}_2)\sqrt{N-2}}{\sqrt{4(1-\hat{\alpha}_1)(1-\hat{\alpha}_2)(1-r_{12})}} = \frac{(0,583-0,718)\sqrt{8-2}}{\sqrt{4(1-0,583)(1-0,718)(1-0,98)}} = -3,41$$

Valors crítics de la distribució t de Student amb 6 ($N-2$) graus de llibertat, per a un nivell de confiança del 95% i contrast bilateral són:

$$t_{0,975(6)} = 2,447 \text{ i } t_{0,025(6)} = -2,447$$

Com que l'estadístic de contrast obtingut ($-3,41$) queda fora de l'interval entre els valors crítics (entre $-2,447$ i $2,447$), podem rebutjar la hipòtesi nul·la i acceptar l'alternativa, i concloure que amb un nivell de confiança del 95% la diferència entre els dos coeficients alfa és estadísticament significativa.

Com en el cas de dues mostres independents, Woodruff i Feldt (1986) també van fer extensiu el contrast anterior a més de dos coeficients comparats simultàniament en mostres dependents. Es pot consultar el seu desenvolupament en els mateixos textos citats per a mostres independents.

4.2.3. Kuder-Richardson

Uns quants anys abans que Cronbach proposés el coeficient alfa com a indicador de la consistència interna d'un test, Kuder i Richardson (1937) van presentar dues fórmules de càlcul d'aquest indicador que de fet són casos particulars d' α quan els ítems són dicotòmics. Aquestes dues fórmules són conegudes com a KR_{20} i KR_{21} .

Quan els ítems d'un test són dicotòmics i es codifiquen les dues alternatives de resposta possibles com a 0 i 1, la variància d'un ítem és igual a la proporció de zeros per la proporció d'uns. Si el test és de rendiment i les respostes als diferents ítems són correctes o incorrectes, habitualment es codifiquen amb un 1 les respostes correctes i amb un 0 les incorrectes. En aquest cas, la variància de l'ítem serà igual a la proporció de subjectes que encerten l'ítem (p_j) per la proporció de subjectes que no l'encerten (q_j). De la mateixa manera, si el test és de personalitat i no hi ha respostes correctes ni incorrectes, però es codifiquen amb un 1 els subjectes que responen *sí* i amb un 0 els que responen *no*, la variància de l'ítem serà la proporció de subjectes que responen *sí* (p_j) per la proporció de subjectes que responen *no* (q_j). En tots dos casos $S_j^2 = p_j q_j$.

Tenint en compte aquesta igualtat, la fórmula del KR_{20} simplement substitueix, en la del coeficient alfa de Cronbach, el sumatori de les variàncies dels ítems pel sumatori dels productes p_j per q_j :

$$KR_{20} = \frac{n}{n-1} \left(1 - \frac{\sum_{j=1}^n p_j q_j}{S_x^2} \right)$$

En cas que tots els ítems tinguessin la mateixa dificultat, o que el nombre de subjectes que responen *sí* es mantingués constant per a tots els ítems, el producte p_j per q_j seria igual per a tots ells, i el seu sumatori seria igual a la mitjana del test menys aquesta mitjana al quadrat dividida pel nombre d'ítems (n). Aquesta nova igualtat permet reformular el KR_{20} per al cas en què tots els ítems tinguin la mateixa dificultat o el nombre de subjectes que responen *sí* es mantingui constant per a tots els ítems. Aquesta reformulació és el KR_{21} .

$$KR_{21} = \frac{n}{n-1} \left(1 - \frac{\bar{X} - \bar{X}^2/n}{S_x^2} \right)$$

Si apliquem aquestes fórmules al nostre exemple, obtenim els resultats següents:

Taula 9

Subjectes	Ítems						x
	1	2	3	4	5	6	
A	1	1	1	1	0	1	5
B	0	1	1	1	1	0	4
C	1	1	0	1	1	0	4
D	1	1	1	1	1	1	6
E	1	1	1	1	1	1	6
F	0	1	1	0	0	0	2
G	0	1	1	0	1	0	3
H	1	0	1	0	0	0	2
$p_j q_j$	0,234375	0,109375	0,109375	0,234375	0,234375	0,234375	

$$\sum p_j q_j = 0,234375 + 0,109375 + 0,109375 + 0,234375 + 0,234375 + 0,234375 = 1,15625$$

$$S_x^2 = 2,25$$

$$\bar{X} = 4$$

$$KR_{20} = \frac{n}{n-1} \left(1 - \frac{\sum_{j=1}^n p_j q_j}{S_x^2} \right) = \frac{6}{5} \left(1 - \frac{1,15625}{2,25} \right) = 0,583$$

$$KR_{21} = \frac{n}{n-1} \left(1 - \frac{\overline{X} - \bar{x}^2 / n}{S_x^2} \right) = \frac{6}{5} \left(1 - \frac{4 - 4^2 / 6}{2,25} \right) = 0,489$$

Com podem comprovar, el KR_{20} proporciona un resultat idèntic al del coeficient alfa, mentre que el KR_{21} dóna un resultat inferior, ja que en el nostre exemple no tots els ítems tenen la mateixa dificultat. En aquest sentit, el càlcul del KR_{21} no seria adequat per a aquest cas.

5. Factors que afecten la fiabilitat

La fiabilitat d'un test depèn de factors com ara la variabilitat de les puntuacions del test, el nombre total d'ítems del test o bé les característiques dels ítems que el componen, entre altres. A continuació, es veuran aquests tres aspectes: variabilitat, longitud i característiques dels ítems.

a) Variabilitat

En els apartats previs s'ha abordat la fiabilitat a partir del càlcul del coeficient de correlació entre dos tests paral·lels, entre dues administracions del test en dos moments temporals diferents o entre diferents parts del test. Cal tenir en compte, però, que el coeficient de correlació és sensible al rang i variabilitat de les dades. El que s'observa és que quan mantenim la resta de factors constants, si s'augmenta la variabilitat de les dades, el coeficient de correlació augmenta. Per la qual cosa, en els casos en què hi hagi una alta variabilitat en les puntuacions del test, el coeficient de fiabilitat serà més alt. D'això es desprèn que un test no té un coeficient de fiabilitat únic i fix, sinó que depèn de les característiques de la mostra sobre la qual es calcula. Així, per contra, si la mostra és homogènia i les puntuacions empíriques que s'obtenen presenten una variabilitat baixa, el coeficient de fiabilitat serà més baix.

En aquest punt, val la pena recordar les paraules de Crocker i Algina (1986) quan diuen que no es pot afirmar que un test és fiable o no ho és, sinó que la fiabilitat és una propietat de les puntuacions obtingudes en el test a partir d'una mostra particular d'individus.

b) Longitud

Un altre dels factors que afecten la fiabilitat és la longitud del test. Així, la fiabilitat depèn del nombre d'ítems que presenti el test. La lògica d'aquesta afirmació resideix en el fet que com més ítems s'utilitzin per a mesurar un constructe, millor podrà ser valorat aquest i més petit serà l'error de mesura que es cometrà en estimar la puntuació vertadera del subjecte. Per la qual cosa, sempre que s'augmenti el nombre d'ítems d'un test, (sempre que aquests siguin ítems representatius del constructe) la fiabilitat augmentarà. Per a conèixer la fiabilitat d'un test en cas que s'augmentin o disminueixin el nombre d'ítems, s'utilitza la fórmula de Spearman Brown, també coneguda com a *profecia de Spearman Brown*.

$$R_{xx} = \frac{k r_{xx}}{1 + (k - 1) r_{xx}}$$

D'aquesta manera, k estarà determinat pel quocient entre el nombre d'ítems finals (n_f) del test dividit pel nombre d'ítems inicials (n_i) del test:

$$k = \frac{n_f}{n_i}$$

Si s'afegeixen ítems a un test, k sempre serà superior a 1, mentre que si s'escurça el test (eliminem ítems) k serà inferior a 1.

Exemple

Si un test de 25 ítems presenta una fiabilitat de 0,65 i hi afegim 10 ítems paral·lels. Quina serà la seva fiabilitat?

$$k = \frac{35}{25} = 1,4$$

k indica que cal allargar 1,4 vegades la longitud del test. Si se substitueix aquest valor en la fórmula:

$$R_{xx} = \frac{1,4 \cdot 0,65}{1 + (1,4 - 1) \cdot 0,65} = 0,72$$

S'obté que el nou coeficient de fiabilitat del test allargat serà de 0,72.

També la pregunta anterior es pot invertir i plantejar-se quants ítems hauria de tenir el test per a assolir una determinada fiabilitat. En aquest cas, caldria aïllar k de la fórmula:

$$k = \frac{R_{xx}(1 - r_{xx})}{r_{xx}(1 - R_{xx})}$$

Efectivament, si ara la pregunta fos quants ítems cal afegir per a aconseguir una fiabilitat de 0,72 a un test de 25 ítems que presenta una fiabilitat de 0,65, s'aplicaria la fórmula anterior per a saber quantes vegades caldria augmentar el test:

$$k = \frac{0,72 \cdot (1 - 0,65)}{0,65 \cdot (1 - 0,72)} = 1,4$$

La qual cosa permetrà saber el nombre d'ítems que cal afegir:

$$k \cdot n_i - n_i$$

$$1,4 \cdot 25 - 25 = 10$$

El test final tindria 35 ítems ($1,4 \times 25 = 35$), per la qual cosa caldria afegir 10 ítems als 25 inicials per a aconseguir una fiabilitat de 0,72.

Cal tenir present dos aspectes, el primer és que malgrat que la fiabilitat d'un test augmentarà sempre que augmentem el nombre d'ítems, aquest augment no és directament proporcional. Per a veure l'efecte de l'augment que es produeix en la fiabilitat en diferents valors de k , cal fixar-se en la figura 1, en

Lectura de la fórmula

R_{xx} : nou coeficient de fiabilitat del test allargat o escurçat.

r_{xx} : coeficient de fiabilitat del test original.

k : nombre de vegades que s'allarga o s'escurça el test.

què en les abscisses es representen els diferents valors de k (el nombre de vegades que s'ha allargat el test) i en les ordenades el coeficient de fiabilitat que s'obtingria.

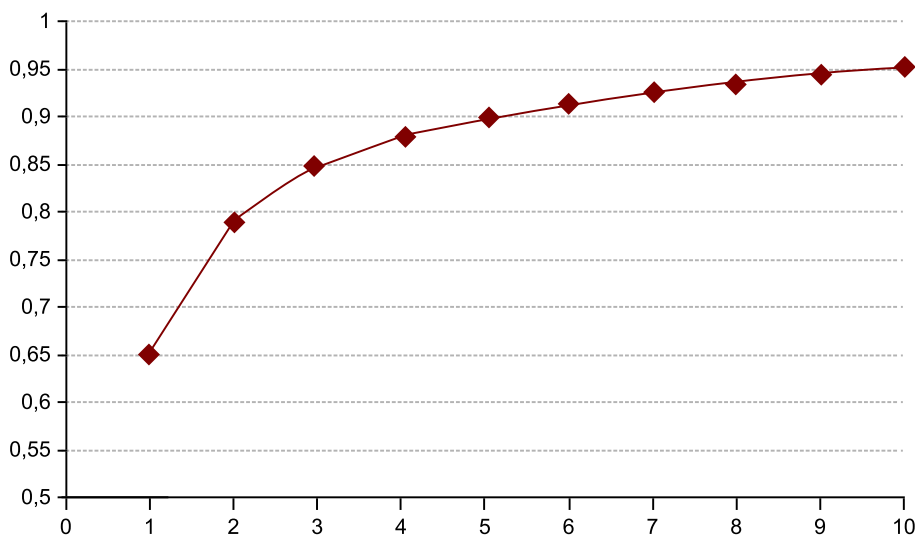


Figura 1. Relació entre el coeficient de fiabilitat i l'augment d'ítems en un test

El segon aspecte que cal tenir en compte és que malgrat que sempre podem aconseguir un augment de la fiabilitat augmentant el nombre d'ítems del test, s'han de valorar els aspectes de fatiga que comporta respondre a un instrument amb molts ítems. Per la qual cosa, es pot pensar a augmentar el nombre d'ítems si el test original té relativament pocs ítems i una fiabilitat que no arriba a ser del tot adequada. Però si el test ja presenta un nombre considerable d'ítems o bé la seva fiabilitat dista molt de ser adequada, potser caldria seleccionar un altre test per a mesurar el constructe o bé construir un nou test.

c) Característiques dels ítems

Cada ítem del test contribueix de manera específica a la fiabilitat o consistència interna del test. Una manera de comprovar-ho és calcular el coeficient alfa de Cronbach eliminant del càlcul la puntuació de l'ítem. Si l'ítem contribueix de manera positiva a la consistència interna del test, en eliminar-lo del test, el valor del coeficient alfa de Cronbach s'alterarà a la baixa (si eliminem l'ítem, el test perd consistència interna). Al contrari, si observem que en eliminar l'ítem el coeficient alfa de Cronbach augmenta, indicarà que l'ítem no contribueix de manera positiva a la consistència interna.

Per exemple, si un test amb 6 ítems presenta una consistència interna de 0,76 i es calcula de nou la consistència interna eliminant la puntuació de l'ítem en cada cas, caldria valorar que en tots els casos la consistència interna fos inferior a la del conjunt del test.

En la taula següent es mostra un exemple:

Taula 10

Ítems	Alfa de Cronbach sense l'ítem
Ítem 1	0,72
Ítem 2	0,74
Ítem 3	0,80
Ítem 4	0,71
Ítem 5	0,70
Ítem 6	0,75

En la taula anterior, s'observa que quan s'elimina un ítem el valor del coeficient alfa de Cronbach disminueix (ítems 1, 2, 4, 5 i 6), la qual cosa indica que l'ítem contribueix de manera favorable a la consistència interna del test. No obstant això, quan s'elimina l'ítem 3, el valor del coeficient alfa de Cronbach augmenta, la qual cosa indica que el test presentaria una consistència interna millor sense aquest ítem.

6. Estimació de la puntuació vertadera

Recordem que, segons la teoria clàssica, la puntuació empírica del subjecte és igual a la puntuació vertadera més l'error aleatori de mesura. Conèixer la precisió amb la qual es mesura el constructe permet calcular la quantitat d'error que està afectant la puntuació empírica. D'aquesta manera, conèixer la fiabilitat de l'instrument permet estimar la puntuació vertadera del subjecte. Tanmateix, cal tenir en compte que no es pot calcular exactament la puntuació vertadera del subjecte, però sí estimar-la a partir del càlcul d'un interval de confiança.

Fonamentalment, s'utilitzen dos procediments per a estimar aquesta puntuació vertadera:

- L'estimació que assumeix la distribució normal de l'error aleatori.
- L'estimació a partir del model de regressió lineal.

6.1. Estimació de la puntuació vertadera a partir de la distribució normal de l'error aleatori

Aquest procediment assumeix que l'error es distribueix segons la llei normal amb mitjana 0 i variància S_e^2 . A partir del càlcul d'un interval de confiança s'obtenen els límits inferior i superior entre els quals es trobarà la puntuació vertadera (V) del subjecte amb un determinat nivell de confiança (o bé assumint un determinat risc alfa). Els passos que caldria seguir són els següents:

1. Calcular l'error típic de mesura (S_e) que està determinat per l'expressió següent:

$$S_e = S_x \sqrt{1 - r_{xx}}$$

2. Cercar el valor $Z_{\alpha/2}$ per al nivell de confiança amb el qual es vulgui treballar. Si s'ha fixat el nivell de confiança al 95% (o bé el risc d'error al 5%), el valor $Z_{\alpha/2}$ que correspon segons les taules de la distribució normal és de 1,96.

3. Calcular l'error màxim de mesura ($E_{màx}$) que s'està disposat a assumir:

$$E_{màx} = Z_{\alpha/2} \cdot S_e$$

4. Calcular l'interval de confiança en què es trobarà la puntuació vertadera del subjecte a partir de l'expressió següent:

Lectura de la fórmula

S_x : desviació típica de les puntuacions del test.
 r_{xx} : coeficient de fiabilitat obtingut.

$$IC = X \pm E_{m\grave{a}x}$$

Exemple

S'ha administrat un test a una mostra de 300 subjectes. La puntuació total del test presenta una mitjana de 15,6 punts, una desviació típica de 5,4 i un coeficient de fiabilitat de 0,77. Si es vol treballar amb un nivell de confiança del 95%, entre quins valors se situaria la puntuació vertadera d'un subjecte que ha obtingut una puntuació empírica de 18?

Error típic de mesura: $S_e = 5,4\sqrt{1 - 0,77} = 2,59$

Un nivell de confiança del 95% correspon a una $z_{\alpha/2} = 1,96$

Error màxim:

$$E_{m\grave{a}x} = 1,96 \cdot 2,59 = 5,08$$

$$IC = 18 \pm 5,08$$

$$12,92 \leq V \leq 23,08$$

La puntuació vertadera del subjecte oscil·larà entre els valors 12,92 i 23,08 amb un nivell de confiança del 95%.

Cal tenir en compte que la precisió de la mesura, és a dir, la fiabilitat de l'instrument que utilitzem per a mesurar el constructe, té un efecte directe en l'amplitud de l'interval construït. Quan utilitzem un instrument amb una fiabilitat alta, l'estimació de la puntuació vertadera del subjecte serà més precisa, cosa que implica que l'amplitud de l'interval serà menor. Al contrari, quan s'utilitzen instruments menys fiables, la precisió amb la qual podem calcular la puntuació vertadera del subjecte és més baixa i l'interval construït és més ampli.

6.2. Estimació de la puntuació vertadera a partir del model de regressió lineal

La puntuació empírica és una puntuació esbiaixada (des d'un punt de vista teòric s'observa una correlació positiva entre la puntuació empírica i l'error de mesura), per la qual cosa l'estimació que es pugui fer a partir d'aquesta també presentarà un biaix. És per això que una altra possibilitat és estimar primer la puntuació vertadera i posteriorment calcular l'interval de confiança a partir de la puntuació vertadera obtinguda.

Si ens basem en el supòsit que la puntuació vertadera és igual a la mitjana de la puntuació empírica podem fer una estimació de la puntuació vertadera a partir de l'equació de regressió següent:

$$V' = r_{xx} (X - \bar{X}) + \bar{X}$$

Una vegada estimada la puntuació vertadera es passaria a calcular l'interval de confiança per a establir amb un determinat nivell de confiança entre quins valors se situaria la puntuació vertadera del subjecte. En aquest cas s'utilitzarà l'error típic d'estimació, desviació típica de la diferència entre la puntuació vertadera i la puntuació vertadera pronosticada ($V - V'$):

$$S_{VX} = S_X \sqrt{1 - r_{xx}} \sqrt{r_{xx}} = S_e \sqrt{r_{xx}}$$

Continuant amb l'exemple anterior, a continuació es mostra com es calcula la puntuació vertadera pronosticada i posteriorment l'interval de confiança que indicarà entre quins valors se situarà la puntuació vertadera del subjecte amb un nivell de confiança del 95%.

En primer lloc es calcula la puntuació vertadera:

$$V' = 0,77 \cdot (18 - 15,6) + 15,6 = 17,45$$

S'obté l'error típic d'estimació a partir de l'error típic de mesura calculat en l'exemple anterior i el coeficient de fiabilitat:

$$S_{VX} = 2,59 \sqrt{0,77} = 2,27$$

A continuació es calcula l'error màxim:

$$E_{m\grave{a}x} = 1,96 \cdot 2,27 = 4,45$$

L'interval de confiança quedarà construït a partir de l'estimació de la puntuació vertadera i l'error màxim obtingut.

$$IC = 17,45 \pm 4,45$$

$$13,00 \leq V \leq 21,90$$

El resultat de l'interval, tant a partir de l'estimació basada en la distribució normal dels errors com a partir del model de regressió lineal, s'interpreta de la mateixa manera. En tots dos casos es pot afirmar que hi ha una probabilitat del 95% que un subjecte que hagi tingut una puntuació empírica de 18 punts en el test tingui una puntuació vertadera situada entre 13 i 23 (o entre 13 i 22 si l'estimació s'ha fet a partir del model de regressió lineal).

Lectura de la fórmula

V' : puntuació vertadera pronosticada.

r_{xx} : coeficient de fiabilitat del test.

X : puntuació empírica obtinguda pel subjecte.

$\bar{X}$: mitjana de les puntuacions del test.

Lectura de la fórmula

S_X : desviació típica de les puntuacions del test.

r_{xx} : coeficient de fiabilitat del test.

S_e : error típic de mesura.

7. Fiabilitat dels tests referits al criteri

7.1. Conceptes bàsics

Els tests als quals s'ha fet referència fins ara són els anomenats *tests referits a la norma* (TRN). Tal com s'ha exposat en el mòdul "Aproximació històrica i conceptes bàsics de la psicometria", aquests tests tenen com a objectiu escalar els subjectes en funció de les puntuacions obtingudes en un test que mesura un determinat tret o variable psicològica. En aquest apartat ens centrarem en un altre tipus de tests, els tests referits al criteri (TRC).

Aquests tipus de tests comencen a fer-se populars cap a la dècada dels setanta del segle XX i a partir d'aquest moment aconseguixen fer-se un lloc dins de la teoria dels tests per a avaluar els estàndards de rendiment. El seu sorgiment s'explica, d'una banda, pel context contrari que s'havia desenvolupat, fonamentalment als Estats Units, sobre l'ús dels tests i, d'una altra, per la necessitat de valorar l'eficàcia dels programes educatius que durant els darrers anys havien sorgit. En aquest darrer punt, calia avaluar el nivell d'habilitat i coneixement que els individus havien assolit dins del programa educatiu per a demostrar-ne l'eficàcia. Actualment, el seu ús està molt estès en el camp educatiu i laboral, en què el que es persegueix és avaluar la competència que presenten els individus en un determinat conjunt de coneixements o criteri.

Un TRC cerca avaluar en termes absoluts l'estatus d'un subjecte respecte a un criteri. Entenent com a *criteri* un domini de conductes, continguts o conjunt de processos ben definits. Per a assolir aquest punt, en primer lloc és necessari definir el conjunt de tasques que l'individu ha de ser capaç de fer per a demostrar la seva competència sobre el criteri. El test es construeix a partir d'una sèrie d'ítems que representen el domini que vol ser avaluat. L'estimació del criteri es fa a partir de la puntuació que l'individu obté en el test. La interpretació d'aquesta puntuació és bastant intuïtiva, ja que la proporció d'ítems correctament encertats indicarà la competència que el subjecte té del criteri. Per exemple, si es vol valorar la capacitat en matemàtiques dels alumnes de tercer de primària, caldrà definir el domini de conductes (de continguts o de processos) que els alumnes haurien de poder posar en pràctica. Una vegada decidits els ítems (preguntes, problemes, exercicis, etc.) que formaran part del test i que representen el criteri que es vol valorar, s'administren als subjectes i la puntuació que s'obtingui serà l'estimació del criteri.

En aquest punt val la pena aturar-se a valorar les diferències entre els TRN i els TRC, ja que, com es mostrarà en els apartats següents, en la pràctica la distinció no està sempre clara. Mentre que els TRN tenen com a finalitat respondre a la

pregunta: la puntuació del subjecte *A* és superior a la del subjecte *B* quan es valora *X*?. Els TRC es plantegen la pregunta en termes diferents: els subjectes *A* i *B* són capaços de resoldre *X*?

En la taula següent es resumeixen les característiques bàsiques d'aquests dos tipus de test respecte a l'objecte que s'avalua, l'anàlisi de la fiabilitat que duen a terme, l'objectiu de l'avaluació i l'àmbit d'aplicació.

Taula 11. Característiques bàsiques dels tests referits a la norma i dels tests referits al criteri

	TRN	TRC
Objecte que s'avalua	Mesura una variable psicològica o tret.	Mesura un conjunt de coneixements o competències (criteri).
Anàlisi de la fiabilitat	A partir de les diferències entre els subjectes.	A partir del domini que presenta l'individu sobre el tema.
Objectiu d'avaluació	Posició relativa d'un individu respecte a la resta del grup.	Grau de coneixement que presenten els individus en el domini.
Àmbit d'aplicació	Personalitat, actituds, etc.	Educació, àmbit laboral, avaluació de programes, etc.

En essència, el concepte de *fiabilitat* en el cas dels TRC i els TRN és idèntic: en tots dos casos es vol valorar el grau d'error que es comet a l'hora de fer un mesurament. No obstant això, els procediments que se segueixen per a estimar la fiabilitat en la teoria clàssica per als TRN no són, en general, apropiats per a estimar la fiabilitat dels TRC. La raó és que, mentre que els TRN basen la mesura de fiabilitat en la variabilitat de les puntuacions del test (recordem que el coeficient de fiabilitat s'ha definit com la proporció de variància de les puntuacions veritables que hi ha en la variància de les puntuacions empíriques), en els TRC aquesta variabilitat deixa de tenir importància. La finalitat que persegueixen els TRC és avaluar el nivell de coneixement que els individus tenen sobre un criteri, i majoritàriament basen la mesura de fiabilitat en la classificació dels subjectes en dues grans categories: els que dominen el criteri i els que no el dominen. Si la classificació que es fa dels individus a partir d'una o més administracions del test és consistent, es pot parlar de *consistència* o *precisió en el procés de mesura* i consegüentment de *fiabilitat*. No obstant això, a diferència dels TRN, els TRC se centren en la idea que la puntuació del test permet fer una interpretació en termes absoluts de la capacitat del subjecte sense haver de comparar els resultats amb un grup de referència. Les diferències entre els individus (la variabilitat de la mesura) deixen de tenir importància per a avaluar la fiabilitat del test (les diferències en la puntuació del test entre uns i altres poden ser petites sense que això afecti la fiabilitat).

Com s'ha vist abans, un dels factors que afecten els valors de la fiabilitat és la variabilitat de les puntuacions del test. Per això els procediments que fins ara s'han presentat per a valorar la fiabilitat no són adequats per a ser aplicats en els TRC. Seguint Shrock i Coscarelli (2007) hi ha tres maneres d'abordar la fiabilitat en aquest tipus de test: els procediments que requereixen dues apli-

cacions del test per a valorar la consistència de la classificació, els que només requereixen una única aplicació i aquells en què entra en joc el paper dels avaluadors. En aquest darrer cas, són els mateixos avaluadors els que jutgen la competència dels subjectes. En els apartats següents es veuran aquests diferents procediments per a abordar la fiabilitat en els TRC.

7.2. Índexs d'acord que requereixen dues aplicacions del test

Es considera que un test és fiable quan s'administra el mateix test dues vegades a la mateixa mostra o bé s'apliquen dues formes paral·leles del test i els resultats de les dues administracions permeten classificar els subjectes dins de la mateixa categoria. A continuació es presenten dos dels índexs d'acord més utilitzats: el coeficient de Hambleton i Novick (p_{H-N}) i el coeficient kappa (k).

7.2.1. Coeficient de Hambleton i Novick

El coeficient de Hambleton i Novick (1973) valora la fiabilitat o consistència de les classificacions a partir de dues administracions del test o de dues formes paral·leles del test. Per a calcular-lo es té en compte, d'una banda, la proporció de subjectes que són classificats de manera consistent en una mateixa categoria (p_c) tant si aquesta és la de competents com si és la de no competents (aptes o no aptes, etc.) i, de l'altra, la proporció de classificacions consistents que s'esperaria per atzar (p_a). La diferència entre la proporció de valoracions consistents i les que s'esperarien per atzar dóna el coeficient de Hambleton i Novick (p_{H-N}):

$$p_{H-N} = p_c - p_a$$

El càlcul de la proporció de subjectes classificats de manera consistent s'expressa de la manera següent:

$$p_c = \sum p_i = \frac{n_{11}}{N} + \frac{n_{00}}{N}$$

Malgrat que l'expressió que es proporciona contempla classificar els individus només en dues categories (competents i no competents), la fórmula es pot aplicar independentment del nombre de categories. En aquest cas caldrà sumar les proporcions de cada una de les categories establertes.

Quan tots els subjectes són classificats de manera consistent en les dues administracions del test, p_c pren el valor màxim d'1. El seu valor mínim està determinat per la proporció de classificacions consistents que s'esperaria per atzar (p_a). Per a calcular aquesta proporció de classificacions consistents que s'esperarien per atzar s'utilitzen les freqüències marginals d'una taula de contingència aplicant la fórmula següent:

Lectura de la fórmula

p_c : proporció de subjectes classificats consistentment en les dues administracions del test.

N : nombre total de subjectes avaluats.

n_{11} : nombre de subjectes classificats com a competents en les dues administracions del test.

n_{00} : nombre de subjectes classificats com a no competents en les dues administracions del test.

$$p_a = \sum \frac{n_{.j} \cdot n_{i.}}{N^2}$$

Lectura de la fórmula

$n_{.j}$: nombre de subjectes classificats com a competents (o no competents) pel test A.

$n_{i.}$: nombre de subjectes classificats com a competents (o no competents) pel test B.

Exemple

Suposem que s'administren dos tests paral·lels de 20 ítems a un grup de 16 individus. Perquè l'individu sigui classificat en el grup de subjectes competents cal que contesti correctament 14 ítems.

En la taula següent es mostren les dades d'aquest exemple. En aquesta taula es representen les puntuacions obtingudes pels 16 individus en els dos tests (A i B) i la classificació que a partir del punt de tall estableix cada un d'aquests tests (classificació del test A i classificació del test B).

Taula 12. Dades d'exemple. Coeficient de Hambleton i Novick

Sub- jecte	Test A	Test B	Classificació test A	Classificació test B
1	10	12	No competent	No competent
2	16	18	Competent	Competent
3	19	20	Competent	Competent
4	8	10	No competent	No competent
5	10	12	No competent	No competent
6	15	16	Competent	Competent
7	14	12	Competent	No competent
8	17	18	Competent	Competent
9	18	19	Competent	Competent
10	13	14	No competent	Competent
11	16	17	Competent	Competent
12	19	17	Competent	Competent
13	12	15	No competent	Competent
14	10	12	No competent	No competent
15	16	14	Competent	Competent
16	18	15	Competent	Competent

A partir de la taula anterior, es construeix la taula de contingència, que veurem a continuació.

Taula 13. Taula de contingència. Coeficient de Hambleton i Novick

		Test B		
Test A		Competents	No competents	
	Competents	9 (n_{11})	1	10 ($n_{1.}$)
	No competents	2	4 (n_{00})	6 (n_{i2})
		11 (n_{1j})	5 (n_{2j})	16 (N)

A partir de la taula de contingència podem calcular el coeficient de Hambleton i Novic:

$$p_c = \frac{9}{16} + \frac{4}{16} = 0,81$$

$$p_a = \frac{10 \cdot 11}{16^2} + \frac{6 \cdot 5}{16^2} = 0,546 \approx 0,55$$

$$p_{H-N} = 0,81 - 0,55 = 0,26$$

Aquest resultat indica que l'ús dels tests permet millorar en un 26% la classificació dels subjectes de la que es faria per atzar.

7.2.2. Coeficient kappa de Cohen

El coeficient kappa (Cohen, 1960) permet estudiar el nivell de concordança en les classificacions a partir de dues administracions del test. Possiblement aquest sigui el coeficient de consistència més extensament utilitzat en la literatura. La seva fórmula és la següent:

$$k = \frac{p_c - p_a}{1 - p_a}$$

En què p_c i p_a són respectivament la proporció de subjectes classificats de manera consistent i la que s'esperaria per atzar tal com s'ha definit abans. Valors propers a 1 indiquen que la consistència en la classificació dels subjectes a partir del test és perfecta, mentre que valors propers a 0 indiquen que la consistència en la classificació és deguda a l'atzar (en aquest cas, l'aplicació dels tests no ha millorat la consistència que es podria obtenir per atzar). En general, els valors del coeficient kappa que oscil·len entre 0,6 i 0,8 es consideren acceptables, i els que se situen per sobre de 0,8 s'interpreten com a molt bons (Landis i Koch, 1977).

A partir de l'ús de l'error típic de mesura (S_e), proposat també per Cohen (1960) pot obtenir-se l'interval de confiança i valorar-ne la significació estadística. L'error típic de mesura i l'interval de confiança es defineixen a partir de les expressions següents:

$$S_{e(k)} = \sqrt{\frac{p_c(1-p_c)}{N(1-p_a)^2}}$$

$$IC = k \pm Z_{\alpha/2} \cdot S_{e(k)}$$

A partir de les dades exposades en l'exemple anterior, el resultat del coeficient kappa seria:

$$k = \frac{0,81 - 0,55}{1 - 0,55} = 0,58$$

Aquest valor indica la consistència en la classificació dels subjectes independentment de la proporció esperada per l'atzar. A partir de l'error típic de mesura s'obté l'interval de confiança amb un nivell de confiança del 95%:

$$S_{e(k)} = \sqrt{\frac{0,81 \cdot (1 - 0,81)}{16 \cdot (1 - 0,55)^2}} = 0,22$$

$$IC_{95\%} = 0,58 \pm 1,96 \cdot 0,22$$

$$0,15 \leq k \leq 1$$

Cal assenyalar que, a causa de la reduïda grandària de mostra, l'interval construït és excessivament ampli. Si en comptes de tenir 16 subjectes fossin 200 els que s'haguessin avaluat s'aconseguiria un interval bastant més precís:

$$S_{e(k)} = \sqrt{\frac{0,81 \cdot (1 - 0,81)}{200 \cdot (1 - 0,55)^2}} = 0,06$$

$$IC_{95\%} = 0,58 \pm 1,96 \cdot 0,06$$

$$0,46 \leq k \leq 0,70$$

7.2.3. Coeficient de Livingston

Els dos procediments anteriors consideren que un error en la classificació dels individus és igual de greu independentment que la puntuació del subjecte se situï a prop o lluny del punt de tall. No obstant això, la lògica ens diu que si un individu ha obtingut una puntuació molt distanciada del punt de tall (és a dir, mostra una competència sobradament alta o bé molt inferior a la del punt de tall) és difícil que en la segona aplicació del test o d'un test paral·lel, el resultat de la classificació sigui contrari a l'obtingut en la primera administració. En aquest sentit, Livingston (1972) proposa un nou coeficient per a avaluar la fiabilitat de les classificacions en el qual té en compte aquest aspecte. Aquest nou coeficient es basa en els mètodes de pèrdua quadràtica, els quals tenen en compte la distància que hi ha entre el punt de tall i la mitjana de les puntuacions del test. El coeficient de Livingston s'expressa a partir de la fórmula següent:

$$K_{xx'}^2 = \frac{r_{xx} S_x S_{x'} + (\bar{x}_x - C)(\bar{x}_{x'} - C)}{\sqrt{[S_x^2 + (\bar{x}_x - C)^2][S_{x'}^2 + (\bar{x}_{x'} - C)^2]}}$$

A continuació es mostra l'aplicació de la fórmula en una situació pràctica.

Suposem que s'han dissenyat dues formes paral·leles d'un examen de psicometria. En aplicar-les a una mostra d'estudiants es va obtenir que la mitjana i la desviació típica del test A va ser de 5,2 i 2,6 respectivament. Mentre que en el test B es va obtenir una mitjana de 5,4 i una desviació típica de 2,3. La correlació entre els dos test va ser de 0,83. El punt de tall per a determinar l'aprobat en la prova es va situar en el 5,5. A partir d'aquestes dades podem calcular $K_{xx'}^2$:

$$K_{xx'}^2 = \frac{0,83 \cdot 2,6 \cdot 2,3 + (5,2 - 5,5) \cdot (5,4 - 5,5)}{\sqrt{[2,6^2 + (5,2 - 5,5)^2] \cdot [2,3^2 + (5,4 - 5,5)^2]}} = 0,83$$

Aquest resultat mostra com la concordança de classificació a partir dels dos tests és bona.

7.3. Índexs d'acord que requereixen una única aplicació del test

El desavantatge principal dels indicadors presentats en l'apartat anterior és que cal aplicar el test dues vegades o bé generar una forma paral·lela del test. Arran d'aquest fet va sorgir la necessitat de trobar algun procediment en què només calgués una única administració del test.

La primera proposta per a avaluar la fiabilitat dels TRC que només requeria una administració del test va venir amb Livingston (1972), que va proposar una lleu modificació a la formulació que s'ha presentat en l'apartat anterior. A partir d'aquesta primera proposta en van seguir moltes altres³. La majoria d'aquests procediments basen el càlcul de la consistència de la classificació utilitzant models estadístics que estimen la puntuació que s'hauria obtingut en una segona administració del test a partir de la puntuació empírica obtinguda pels subjectes. No obstant això, cal dir que la majoria d'aquests procediments són complexos i poc utilitzats en la pràctica professional. Per això en aquest apartat només s'abordarà la proposta inicial de Livingston (1972) per ser una de les més senzilles de càlcul i possiblement la més utilitzada en la pràctica (Shrock i Coscarelli, 2007).

7.3.1. Coeficient de Livingston (una única aplicació)

La proposta que s'ha presentat del coeficient de Livingston en l'apartat anterior es pot modificar fàcilment perquè sigui vàlida quan només s'ha administrat el test una vegada. En aquest cas, cal que es tingui en compte com a coeficient de fiabilitat el coeficient de consistència interna del test. La fórmula proposada per Livingston (1972) és la següent:

Lectura de la fórmula

r_{xx} : coeficient de fiabilitat a partir del procediment de formes paral·leles o tests-retests.

S_x i $S_{x'}$: corresponen respectivament a la desviació típica del test en la primera i segona administració o en cada una de les formes paral·leles del test.

$\bar{x}_x$ i $\bar{x}_{x'}$: corresponen respectivament a la mitjana del test en la primera i segona administració o en cada una de les formes paral·leles del test.

C : punt de tall.

S_x^2 i $S_{x'}^2$: corresponen respectivament a la variància del test en la primera i segona administració o en cada una de les formes paral·leles del test.

⁽³⁾Subkoviak (1976), Huynh (1976), Brennan i Kanne (1977), Breyer i Lewis (1994), Livingston i Lewis (1995), Brennan i Wan (2004), Lee (2005), Lee, Brennan, Wan (2009), entre d'altres.

$$K^2 = \frac{r_{xx} S_x^2 + (\bar{x} - C)^2}{S_x^2 + (\bar{x} - C)^2}$$

Si només tinguéssim els resultats del primer examen de psicometria de l'exemple anterior i el coeficient alfa de Cronbach fos de 0,78, K^2 seria:

$$K^2 = \frac{0,78 \cdot 2,6^2 + (5,2 - 5,5)^2}{2,6^2 + (5,2 - 5,5)^2} = 0,783$$

Cal tenir present que quan la mitjana coincideix amb el punt de tall, K^2 és igual al coeficient de fiabilitat. Per exemple, en el cas anterior, es pot observar que la distància entre la mitjana i el punt de tall no és gaire gran, per la qual cosa el valor de K^2 és molt similar al valor del coeficient alfa de Cronbach. En canvi, s'observa que quan el punt de tall es distancia de la mitjana, K^2 augmenta. Si ara consideréssim que el punt de tall és de 6,5 el resultat seria:

$$K^2 = \frac{0,78 \cdot 2,6^2 + (5,2 - 6,5)^2}{2,6^2 + (5,2 - 6,5)^2} = 0,82$$

També s'observa que quan augmenta el coeficient de fiabilitat del test també augmenta K^2 . Ara, si augmentéssim el coeficient d'alfa de Cronbach a 0,85, veuríem que el resultat de K^2 també augmentaria:

$$K^2 = \frac{0,85 \cdot 2,6^2 + (5,2 - 5,5)^2}{2,6^2 + (5,2 - 5,5)^2} = 0,852$$

Per la qual cosa es demostra que K^2 sempre serà igual o més gran que el coeficient de fiabilitat del test, i que quan el coeficient de fiabilitat sigui 1, K^2 també prendrà el valor màxim d'1.

$$K^2 \geq r_{xx}$$

7.4. Fiabilitat interobservadors

Els procediments presentats en els apartats previs feien referència a la capacitat del test per a poder classificar de manera consistent els subjectes. En aquest apartat es presentaran alguns dels indicadors més freqüents per a poder avaluar la consistència de les avaluacions fetes per diferents jutges. En alguns contextos educatius i d'empresa, és freqüent que la competència d'un individu pugui ser avaluada per diferents observadors. En aquests casos la qualitat de la mesura depèn de la consistència que s'observa entre l'avaluació dels observadors. El coeficient kappa per a dades nominals o ordinals i el coeficient de concordança per a variables contínues són els dos coeficients de correlació més àmpliament utilitzats que es presentaran en aquest apartat.

Lectura de la fórmula

r_{xx} : coeficient de fiabilitat a partir del procediment de les dues meitats, KR_{20} o alfa de Cronbach.

S_x^2 : variància de les puntuacions del test.

$\bar{x}$: mitjana de les puntuacions del test.

C: punt de tall.

Vegeu també

Recordeu que hem vist el coeficient kappa en l'apartat "Coeficient kappa de Cohen" en aquest mateix mòdul.

7.4.1. Coeficient kappa

En aquest cas el coeficient kappa s'aplica quan es vol estudiar la concordança que hi ha entre les valoracions fetes per dos avaluadors. La fórmula, tal com s'ha presentat abans, és la següent:

$$k = \frac{p_c - p_a}{1 - p_a}$$

En què p_c correspon a la proporció de subjectes classificats de manera consistent pels dos (o més) avaluadors, i p_a correspon a la proporció de concordances que s'esperaria trobar entre els dos avaluadors per atzar.

Suposem que dos avaluadors han classificat en dues categories vuitanta treballadors que han seguit un curs de formació: els que han assolit les competències bàsiques per a poder dur a terme les noves tasques (competents), i els que es considera que no (no competents). En la taula de contingència següent es presenten els resultats de les seves valoracions:

Taula 14

		Observador 1		
Observador 2		Competents	No competents	
	Competents	49 (n_{11})	11	60 ($n_{1\cdot}$)
	No competents	1	19 (n_{00})	20 (n_{i2})
		50 (n_{1j})	30 (n_{2j})	80 (N)

$$p_c = \frac{n_{11}}{N} + \frac{n_{00}}{N} = \frac{49}{80} + \frac{19}{80} = 0,85$$

$$p_a = \sum \frac{n_{j\cdot} \cdot n_{\cdot i}}{N^2} = \frac{60 \cdot 50}{80^2} + \frac{30 \cdot 20}{80^2} = 0,523$$

$$k = \frac{p_c - p_a}{1 - p_a} = \frac{0,85 - 0,523}{1 - 0,523} = 0,69$$

Aquest valor s'interpretaria com una consistència acceptable entre els dos avaluadors.

7.4.2. Coeficient de concordança

Lin (1989) va proposar un coeficient de fiabilitat per a calcular el grau d'acord en les valoracions fetes per dos avaluadors quan aquestes fossin de caràcter continu. Malgrat que en apartats previs s'ha vist que quan les variables són contínues s'aplica el coeficient de correlació de Pearson (r_{xy}), en aquest cas no seria adequat aplicar-lo. La raó és que el coeficient de correlació de Pearson tan

sols valora si l'ordre de les valoracions de tots dos avaluadors coincideixen, però no si els valors assignats a aquestes valoracions són realment els mateixos. Per això es podrien obtenir coeficients de correlació de Pearson alts (si l'ordenació entre els dos avaluadors és similar) malgrat que cap valoració no fos coincident. El coeficient de concordança proposat per Lin (1989) permet valorar el grau en què els valors absoluts assignats per cada avaluador són concordants. El coeficient queda definit per l'expressió següent:

$$CC = \frac{2r_{xy}S_xS_y}{S_x^2 + S_y^2 + (\bar{x} - \bar{y})^2}$$

El coeficient pot prendre valors entre 1 i -1, però per a la qüestió que interessa valorar (coincidència de les puntuacions entre els avaluadors) només tindrà sentit quan aquest prengui valors positius. Cal tenir en compte, però, que a l'hora de valorar el resultat, aquest coeficient interpreta el grau d'acord de manera molt més exigent: Valors superiors a 0,99 s'interpreten com una concordança gairebé perfecta, entre 0,95 i 0,99 es parla d'una concordança substancial, entre 0,90 i 0,95 moderada, i per sota de 0,90 es considera que la concordança és pobre.

Exemple

Dos avaluadors corregeixen de manera independent els treballs de 40 alumnes. La correlació de Pearson entre els dos avaluadors és de 0,96. La mitjana de les qualificacions de l'avaluador A és de 7,23 amb una desviació típica de 3,44, mentre que per a l'avaluador B la mitjana és de 6,33 amb desviació típica de 2,88. Quin és el coeficient de concordança entre els dos avaluadors?

$$CC = \frac{2 \cdot 0,96 \cdot 3,44 \cdot 2,88}{3,44^2 + 2,88^2 + (7,23 - 6,33)^2} = 0,91$$

El coeficient de concordança indica que el grau d'acord entre els dos avaluadors és moderat.

Lectura de la fórmula

r_{xy} : valor del coeficient de correlació de Pearson entre les dues valoracions.

S_x i S_y : corresponen respectivament a la desviació típica de l'avaluador 1 i de l'avaluador 2.

S_x^2 i S_y^2 : corresponen respectivament a la variància de l'avaluador 1 i de l'avaluador 2.

$\bar{x}$ i $\bar{y}$: corresponen respectivament a la mitjana de les avaluacions fetes per l'avaluador 1 i 2.

8. Estimació dels punts de tall

Els procediments que s'han presentat fins ara requereixen establir prèviament un punt de tall per a calcular la fiabilitat. Tal com apunta Berk (1986), el punt de tall és el punt que permet prendre decisions i classificar els subjectes com a competents (els que dominen el criteri) i no competents (els que no el dominen). Atès que el punt de tall és una puntuació empírica del test, té associat un error com a puntuació del test, i un error com a puntuació a partir de la qual es prenen decisions. El primer error és l'error de mesura que s'ha presentat en els apartats previs i que s'expressa a partir de l'error típic de mesura. El segon error és un error de classificació i pot prendre dos formats:

- Un individu que és competent i que erròniament és classificat com a no competent (fals no competent).
- Un individu que no és competent i que erròniament és classificat com a competent (fals competent).

Malgrat que és important evitar aquests dos tipus d'errors, hi ha situacions en què cometre un tipus d'error o un altre té repercussions diferents i és més o menys greu.

Exemple

Suposem que després d'un curs de formació en una empresa s'avalua la competència dels treballadors per a poder portar a terme una nova tasca. Una vegada finalitzat el curs, els participants fan un examen per valorar si han assolit les competències bàsiques per a dur a terme la nova tasca. L'empresa decideix que els que no han superat la prova participin en unes noves xerrades per intentar adquirir la competència. En una situació d'aquest tipus es considera que les conseqüències de classificar un subjecte com a competent sense ser-ho és més greu que classificar un subjecte com a incompetent quan en realitat no ho és. En el primer cas, el treballador no sabrà dur a terme la nova tasca, s'equivocarà en fer-la amb la possibilitat que la seva mala praxi provoqui conseqüències més greus. En el segon cas, s'envia un subjecte competent a rebre unes xerrades addicionals que segurament li permetran consolidar els coneixements que ja tenia adquirits.

En aquest apartat abordarem alguns dels mètodes més freqüents per a establir aquest punt de tall. Tal com suggereixen Cizek i Bunch (2007) cada un dels mètodes barreja una part d'art i una part de ciència. Cal tenir en compte, però, que cada un dels mètodes pot portar a un resultat diferent (diferent punt de tall i diferent percentatge de subjectes classificats en cada grup) i que, en definitiva, malgrat que alguns mètodes són més adequats en funció del tipus de tests o circumstància, cap no ha demostrat ser superior als altres. Tots tenen en comú que uns jutges experts determinaran el punt de tall de manera sistemàtica a partir de l'evidència que tenen sobre allò que volen valorar. Segons Jaeger (1989), els diferents mètodes es poden classificar en dos grans grups: d'una banda, els mètodes que se centren en la valoració del test (*test centered*) i, de

l'altra, els que se centren en l'execució dels examinands (*examinee centered*). No obstant això, per motius didàctics, l'exposició que se seguirà dels diferents mètodes es farà sobre la base de tres criteris:

- Mètodes que es basen en la valoració que un grup d'experts o jutges fan sobre els ítems d'un test (*test centered*). Es presentaran els mètodes de Nedelsky (1954), Angoff (1971, 1984), i Sireci, Hambleton i Pitoniak (2004).
- Mètodes que es basen en la valoració d'un grup d'experts sobre la competència dels subjectes (*examinee centered*). Aquí s'exposaran els mètodes del grup de contrast (Berk, 1976) i del grup límit (Zieky i Livingston, 1977).
- Mètodes que es basen en la posició del subjecte respecte al grup normatiu, o també anomenats *mètodes de compromís*, que intenten lligar aspectes referents als TRC i als TRN. En aquest apartat es presentaran els mètodes de Hofstee (1983) i Beuk (1984).

8.1. Mètodes basats en l'avaluació d'experts sobre els ítems

Els mètodes que es presenten a continuació tenen en comú que els participants seran jutges experts sobre el contingut de la matèria que avaluen. En tots els casos se'ls requerirà que valorin si els ítems del test poden ser contestats correctament per determinats individus, sense que per això sigui necessari tenir informació real sobre la competència dels individus als quals s'administrarà la prova. D'aquest últim aspecte es desprèn un dels avantatges principals d'aquests mètodes, i és que es poden aplicar abans que els subjectes contestin el test, ja que no es necessiten dades sobre la seva execució.

8.1.1. Mètode de Nedelsky

El mètode que va proposar Nedelsky (1954) s'aplica sobre ítems de resposta múltiple. S'utilitza encara avui dia de manera àmplia en l'àmbit acadèmic per a determinar el punt de tall i poder determinar si els individus tenen els coneixements mínims sobre una matèria concreta. Un concepte clau d'aquest procediment i d'altres que es veuran en aquest apartat és que els experts o jutges han de tenir en ment un hipotètic grup de subjectes que es trobarien en el límit per a considerar-los competents (o incompetents).

El mètode es basa en el fet que un grup d'experts, sobre el contingut que es vol valorar, determina, per a cada ítem, les alternatives de resposta que un subjecte amb els coneixements mínims requerits sobre la matèria per a superar la prova (amb la competència mínima requerida) rebutjaria com a incorrectes. El recíproc de les alternatives restants ($1 / \text{nombre d'alternatives restants}$) és

el que s'anomena **valor Nedelsky**. Aquest valor s'interpreta com la probabilitat que un individu amb una capacitat mínima sobre la matèria seleccioni l'alternativa correcta.

Exemple

Per il·lustrar el procediment imaginem un ítem amb cinc alternatives de resposta. El jutge determina que dues de les alternatives de resposta poden ser descartades fàcilment per un subjecte que presenti els coneixements mínims requerits sobre la matèria. El subjecte, doncs, té tres alternatives restants. El valor Nedelsky en aquest cas serà $1/3 = 0,33$. Aquest procediment caldria repetir-lo per cada ítem del test i posteriorment caldria sumar tots els valors obtinguts per cada un dels ítems. La suma de tots els valors s'utilitza com a punt de tall del test, el qual servirà per a classificar els individus en les categories de competents i no competents. No obstant això, cal tenir en compte que és habitual que en el procés participin més d'un jutge o avaluador, per la qual cosa caldrà fer la mitjana de les valoracions de cada un dels jutges i posteriorment fer el sumatori de les valoracions mitjanes. En la taula següent es mostra un exemple de càlcul del punt de tall a partir de la valoració de quatre jutges d'un examen de dotze ítems amb cinc alternatives de resposta.

Taula 15. Exemple de càlcul del mètode Nedelsky

	Avaluadors				
	A	B	C	D	
Ítems	Valors Nedelsky				Mitjanes
1	0,5	0,33	0,5	0,5	0,4575
2	1	0,5	1	1	0,875
3	0,2	0,25	0,2	0,25	0,225
4	0,33	0,5	0,25	0,25	0,3325
5	0,33	0,33	0,33	0,25	0,31
6	0,2	0,33	0,5	0,5	0,3825
7	1	0,33	0,5	0,33	0,54
8	0,5	0,25	0,33	0,5	0,395
9	0,25	0,2	0,2	0,2	0,2125
10	0,33	0,33	0,5	0,5	0,415
11	1	1	0,5	0,5	0,75
12	0,33	0,5	0,5	0,5	0,4575
Sumatori					5,3525

En aquest exemple el sumatori de totes les mitjanes és de 5,35, la qual cosa indica que un subjecte amb la competència mínima requerida sobre la matèria per a superar la prova hauria de contestar 5 ítems correctament. Així doncs, es recomanaria que els subjectes amb 5 ítems correctes fossin classificats com a competents, és a dir, que aprovessin l'examen.

La limitació principal del mètode és que els valors que poden prendre les probabilitats assignades als ítems són molt limitats i no presenten intervals iguals entre ells. Per exemple, en el cas anterior els valors de Nedelsky poden ser 0,2, 0,25, 0,33, 0,5 i 1, i com es pot veure les distàncies entre ells no són iguals. Això fa que, en els casos en què s'observa una distància més gran entre els

valors de la probabilitat (per exemple entre els valors 0,5 i 1), molts avaluadors tendeixin a puntuar preferiblement els ítems amb una probabilitat de 0,5 abans que amb una probabilitat d'1 (probabilitat que un subjecte amb la capacitat mínima requerida sobre el contingut sigui capaç de descartar totes les alternatives incorrectes). Aquesta actuació (atribuir una probabilitat de 0,5 en comptes d'1) implica un biaix a la baixa en la puntuació de tall de la prova, la qual cosa fa que aquest mètode proporcioni punts de tall més baixos en comparació a altres mètodes.

8.1.2. Mètode d'Angoff

El mètode d'Angoff (1971) és el més utilitzat en la pràctica i el que presenta més variants respecte al seu plantejament inicial (Cizek i Bunch, 2007). Com en el mètode de Nedelsky, un grup d'avaluadors ha de basar els seus judicis tenint en ment un hipotètic grup d'individus que presenta una competència mínima sobre la matèria per a poder superar una prova. Cal destacar que, a diferència del mètode plantejat per Nedelsky, Angoff (1971) va proposar que els jutges es fixessin en l'ítem globalment i no en cada una de les alternatives de resposta.

La proposta inicial d'Angoff (1971) era que l'avaluador havia de determinar si un individu hipotètic, amb la competència mínima requerida, seria capaç de contestar l'ítem correctament. A l'ítem s'hi assignava un 1, si es considerava que el subjecte era capaç de contestar correctament l'ítem, i es puntuava amb un 0 si es considerava que aquesta persona fallaria l'ítem. El sumatori de les puntuacions assignades a cada ítem seria igual al punt de tall de la prova, el qual permetria distingir els subjectes competents dels que no presentessin una competència mínima requerida sobre la matèria. No obstant això, la versió més estesa d'aquest mètode és el que el mateix Angoff va plantejar en un peu de pàgina en la seva descripció inicial del mètode. En aquest cas proposava una variació del mètode en què, en comptes de pensar en un individu concret, els jutges fessin les seves valoracions tenint en ment un grup d'individus amb una competència mínima per a superar la prova i valoressin la proporció de subjectes que contestarien l'ítem correctament. El sumatori de les probabilitats assignades a cada un dels ítems representaria el punt de tall de la prova. L'exemple següent mostra un cas hipotètic d'aquesta última opció.

Exemple

Se sol·licita a quatre avaluadors que indiquin el percentatge de subjectes que seria capaç de resoldre correctament cada un dels 12 ítems d'un examen de psicometria. S'indica als avaluadors que cal centrar les seves valoracions pensant en un possible grup d'individus que suposadament presenta la competència mínima requerida sobre la matèria. En la taula següent es mostren els resultats d'aquest cas hipotètic.

Taula 16. Exemple de càlcul del mètode d'Angoff

	Avaluadors			
	A	B	C	D
Ítems	Percentatge de subjectes que resolen correctament l'ítem			
1	30	40	45	38
2	80	90	75	85
3	85	70	65	70
4	50	45	50	60
5	60	50	45	55
6	80	90	70	75
7	40	45	50	35
8	20	25	30	10
9	90	80	85	90
10	30	40	35	40
11	40	45	50	50
12	60	65	70	75
Mitjana	55,42	57,08	55,83	56,92

En aquest cas, el punt de tall s'obté a partir del càlcul de la mitjana de les puntuacions atorgades pels quatre jutges:

$$C = \frac{55,42 + 57,08 + 55,83 + 56,92}{4} = 56,31\%$$

Aquest resultat indica que el punt de tall recomanat per a aquest test seria que els subjectes contestessin correctament un 56,31% dels ítems, és a dir, 6,76 ítems (7 ítems) del conjunt de 12 ítems del test.

Un avantatge important d'aquest mètode és que el fet d'avaluar globalment l'ítem permet utilitzar ítems d'altres formats, a més dels ítems d'elecció múltiple utilitzats en el mètode de Nedelsky. Això comporta que el mètode pot ser utilitzat quan els ítems de la prova són en format obert (Hambleton i Plake, 1995). Com s'ha comentat abans, les variants del mètode són moltes, per exemple una molt usada per a evitar opinions divergents entre els jutges és demanar als avaluadors que valorin els ítems en diferents rondes emprant un procediment similar al que s'utilitza en un mètode Delphi.

8.1.3. Mètode del consens directe

El mètode del consens directe és un mètode relativament recent proposat per Sireci, Hambleton i Pitoniak (2004). Alguns dels avantatges destacables respecte als mètodes anteriors són que, d'una banda, cada jutge expert que participa

en el procés pot expressar directament la seva opinió sobre quina hauria de ser la localització concreta del punt de tall i, d'una altra, que permet als jutges modificar les seves valoracions a partir de les opinions d'altres participants.

El procediment que se segueix en aquest mètode requereix que els ítems del test s'agrupen en seccions. Aquestes seccions es formen a partir de subàrees de contingut homogènies. La funció dels jutges consisteix a dir quants ítems de cada secció podrien ser contestats correctament per subjectes que tenen una competència mínima per a poder ser considerats competents. El sumatori dels ítems en cada subàrea que es considera que es contestaran correctament serà igual a la puntuació de tall.

Immediatament després que els jutges han fet les valoracions, es presenten a tot el panel d'experts els resultats de cada un dels membres i es procedeix a discutir les raons de les possibles diferències. En aquesta fase es promou una discussió oberta entre els jutges perquè defensin i raonin el perquè de les seves valoracions. L'objectiu és que la discussió faciliti el consens entre els avaluadors que permeti una convergència en el punt de tall. Una vegada finalitzada la discussió, se'ls ofereix l'oportunitat de tornar a valorar cada una de les seccions del test. Per acabar, una vegada recollides les dades de la segona valoració, en una segona ronda es comenten les diferents puntuacions per a intentar una major convergència en el punt de tall del test.

Exemple

Es va sol·licitar a quatre jutges que valoressin una prova de psicometria de 60 ítems. La prova es va dividir en 4 seccions (anàlisi dels ítems, fiabilitat, validesa i transformació de puntuacions) i cada secció presentava un nombre diferent d'ítems. En la taula següent es presenten per columnes el nombre d'ítems que cada jutge considera que contestarà correctament un subjecte amb coneixements mínims per a superar la prova. En l'última filera es recull el sumatori d'ítems de cada expert, el qual representa el seu punt de tall recomanat per a la prova. En les dues darreres columnes es mostren les mitjanes i desviacions típiques (DT) d'aquestes valoracions i el percentatge d'ítems que serien contestats correctament en cada secció.

Taula 17. Exemple hipotètic del mètode del consens directe

	Avaluadors					
	A	B	C	D		
Seccions	Nombre d'ítems contestats correctament				Mitjana i DT	%
Anàlisi dels ítems (14 ítems)	8	7	8	8	7,75 (0,5)	64,58
Fiabilitat (20 ítems)	14	12	13	10	12,25 (1,71)	68,06
Validesa (16 ítems)	10	11	10	11	10,50 (0,58)	65,63
Transformació de puntuacions (10 ítems)	6	7	8	7	7,00 (0,82)	70
Sumatori	38	37	39	36	37,5	62,50

Les desviacions típiques indiquen la variabilitat d'opinió entre els jutges per cada secció. En l'exemple, els jutges tenen judicis bastant homogenis en valorar els ítems referents a la validesa (malgrat que els ítems que hagi pogut assenyalar cada un dels jutges puguin ser en cada cas diferents), ja que la desviació típica respecte a la seva mitjana és més petita. En canvi, la secció de fiabilitat presenta més desacord entre els jutges, ja que la seva desviació típica respecte a la mitjana és la més alta amb relació a la resta de seccions. L'objectiu de la discussió d'aquests resultats amb el panel de jutges és que es promogui una major convergència en les dades, que ajudi a assolir un consens en el punt de tall que caldria prendre per a aquesta prova.

8.2. Mètodes basats en l'avaluació d'experts sobre la competència dels subjectes

Els mètodes que es presentaran en aquest apartat es caracteritzen perquè els jutges, a més de ser experts en la matèria que avaluen, han de conèixer la competència dels subjectes. Aquest fet comporta que les valoracions que faran no seran sobre subjectes hipotètics, sinó sobre individus reals. Bàsicament és en aquest aspecte en què radica la crítica més ferotge sobre aquests mètodes, ja que a l'hora de fer les valoracions els avaluadors poden estar influenciats, no únicament pels coneixements o habilitats dels subjectes, sinó per altres variables intrínseques a l'individu que poc tenen a veure amb la competència, com ara el sexe, la personalitat, la raça, el comportament, etc. Altres tipus de biaix que s'ha observat a l'hora de fer avaluacions per part de jutges es poden consultar en la revisió de Martínez-Arias (2010).

8.2.1. Mètode del grup de contrast

El mètode del grup de contrast va ser proposat inicialment per Berk (1976). En aquest mètode cal que els jutges classifiquin els subjectes en dos grups en funció del nivell de competència que se suposa que mostraran en la matèria avaluada. És necessari, doncs, que els jutges coneguin sobradament el rendiment que exhibiran els subjectes que han de classificar. En un grup classifiquen els individus que consideren que seran competents i, en l'altre, els que consideren que seran no competents. Una vegada feta aquesta classificació cal administrar la prova als subjectes i puntuar-la. Per a determinar el punt de tall, se solen utilitzar diferents procediments: representar gràficament les puntuacions de la prova de cada un dels grups, utilitzar com a punt de tall algun indicador de tendència central (mitjana o mediana) o bé utilitzar l'anàlisi de regressió logística. No obstant això, l'ús d'aquest últim procediment només es recomana si les mostres són prou grans (Cizek i Bunch, 2007).

Possiblement, determinar el punt de tall a partir de la representació gràfica és el mètode més senzill. Consisteix a representar en una mateixa gràfica les distribucions dels grups: els que han estat classificats com a competents pels jutges i els que han estat valorats com a no competents (vegeu la figura 2). La intersecció entre totes dues distribucions seria el punt de tall de la prova. Idealment se cerca el punt de tall en què tots els subjectes valorats com a competents se situïn per sobre del punt de tall, mentre que els no competents es trobin per sota. En la pràctica, tal com es veu en la figura 2, hi ha una super-

posició de les dues distribucions, amb la qual cosa en funció d'on se situï el punt de tall hi haurà més subjectes competents classificats com a no competents o a l'inrevés.

Per exemple, si observem la figura 2, si es moguéssim la línia vertical (que indicaria el punt de tall) cap a l'esquerra, el nombre de falsos negatius es reduiria, és a dir, es disminuiria la possibilitat de considerar un competent com a no competent (malgrat que s'augmentaria el nombre de subjectes no competents que es consideren competents). Mentre que si la línia es moguéssim cap a la dreta disminuirien els falsos positius, és a dir, els que tot i ser no competents, se'ls considera competents. Així doncs, per a fixar el punt de tall, caldrà valorar en cada cas quin tipus d'error es vol evitar.

Cercar el punt de tall a partir de la figura 2 pot resultar una tasca bastant subjectiva, és per això que alguns autors prefereixen basar la posició del punt de tall en la mitjana o la mediana dels dos grups.

En l'exemple, a partir de les dades que hem representat en la figura 2 suposem que un grup de professors va classificar els seus alumnes de l'assignatura de psicometria en competents i no competents ($n = 258$). Els professors van classificar 90 alumnes com a no competents i 168 com a competents. Els estudiants van fer la prova i la mediana del grup de no competents va ser de 22 punts i la mediana del grup de competents de 44. El punt mitjà entre les dues medianes (en aquest cas 33) pot funcionar com a punt de tall de la prova.

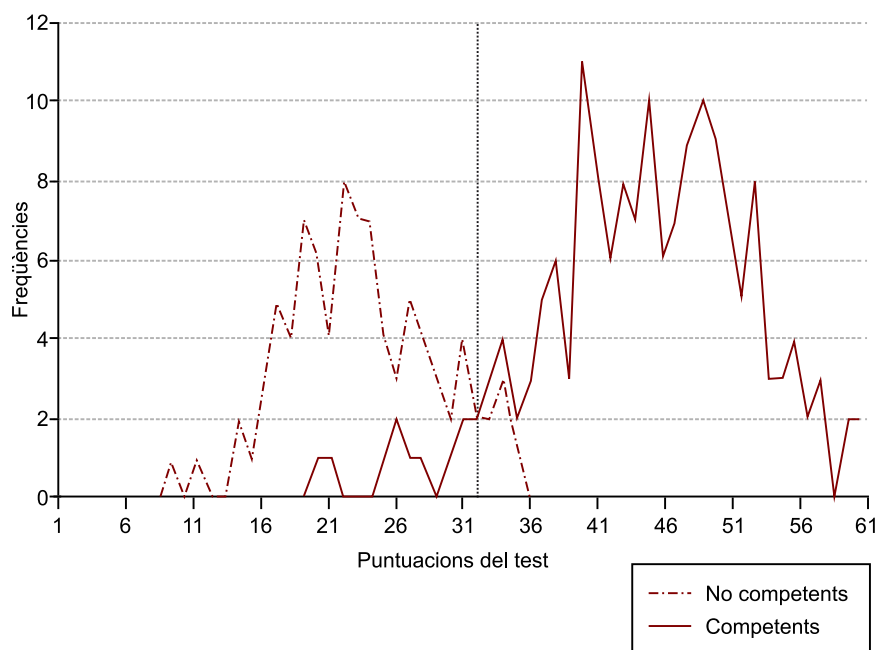


Figura 2. Mètode del grup de contrast. Distribucions del grup de competents i no competents en les puntuacions d'un test

8.2.2. Mètode del grup límit

Sovint pot resultar difícil als jutges classificar els subjectes clarament en dos grups, com a competents o no competents. En aquest sentit, el mètode del grup límit (Zieky i Livingston, 1977) pal·lia aquest aspecte i es pot utilitzar com a alternativa al mètode dels grups de contrast. En aquest mètode, generalment es demana als jutges que classifiquin els subjectes en tres grups: un

grup de subjectes que clarament són competents, un altre grup que clarament es percep com a no competent i, finalment, un altre que se situaria entre els dos grups anteriors. És a dir, un grup límit, en el qual els subjectes tindrien un comportament que els situaria entremig del grup de competents i no competents. Una vegada s'han classificat els subjectes s'administra la prova. Habitualment, el valor de la mediana del grup límit en la prova és el que s'utilitza com a punt de tall. Atesa la senzillesa del mètode, altres autors (Plake i Hambleton, 2001) van proposar una versió alternativa per valorar grups de subjectes amb competència bàsica, notable i avançada.

8.3. Mètodes de compromís

El terme *mètode de compromís* va ser proposat per Hofstee (1983) per a recollir la idea que calien nous procediments que combinessin els coneixements que un subjecte presentava sobre la matèria (valoració de la competència en termes absoluts) però també el nivell d'execució que presentava respecte al seu grup normatiu (valoració de la competència en termes relatius). Els mètodes que s'han presentat fins ara per a fixar el punt de tall són mètodes que es basen exclusivament en la competència del subjecte sobre la matèria que s'ha d'avaluar (valoració en termes absoluts). No obstant això, en la pràctica, els judicis que emeten els avaluadors no es poden basar estrictament en un criteri absolut, sinó que en qualsevol procés d'avaluació es té en compte informació referent al grup normatiu de referència.

Un exemple molt clarificador que il·lustra l'ús combinat d'informació sobre el criteri i sobre el grup normatiu en qualsevol procés d'avaluació el proporcionen Cizek i Bunch (2007). Els autors proposen una situació en què es valora la competència de control d'esfínters d'un nen. Els pares segueixen les rutines habituals perquè el seu infant adquireixi un control adequat, fins que arriba el gran dia en què poden afirmar que el seu fill ha adquirit la competència. Fins aquí tot sembla bastant corrent, però la percepció de la situació canviaria si diguéssim que el nen d'aquest exemple té nou anys? Evidentment, sí.

Aquest exemple demostra que les persones utilitzen múltiples fonts per a valorar les diferents situacions que es troben i que qualsevol intent d'anular o reduir alguna d'aquestes fonts ocasionaria una valoració artificial i poc ajustada a la realitat.

Els mètodes que es presenten a continuació accepten que els tests referits a la norma i al criteri estan, en realitat, bastant units a l'hora de valorar els individus, ja que utilitzen les normes per a poder fixar el criteri. En aquest apartat es presentaran dos dels procediments més habituals: el mètode de Hofstee (1983) i el mètode de Beuk (1984).

8.3.1. Mètode de Hofstee

Hofstee va proposar aquest mètode el 1983 quan després d'impartir la mateixa assignatura durant alguns anys es va trobar que en aquell curs els alumnes presentaven un rendiment molt baix respecte a anys anteriors (sense que cap aspecte significatiu respecte al material, professors, etc. variessin d'un any a l'altre). Després de rebaixar la nota de tall a un 4,5, només el 55% dels estudiants aprovava l'assignatura. Quan en anys anteriors la nota de tall s'havia

situat en un 6 i tot i així aprovava el 90% dels alumnes. El mètode que Hofstee proposa requereix que els avaluadors responguin a quatre preguntes sobre els subjectes que seran avaluats:

- 1) Quin és el punt de tall més alt que es considera acceptable, malgrat que tothom arribi a aquesta puntuació? Se simbolitza per $k_{m\grave{a}x}$.
- 2) Quin és el punt de tall més baix que es considera acceptable, malgrat que ningú arribi a aquesta puntuació? Se simbolitza per $k_{m\grave{i}n}$.
- 3) Quin és el percentatge màxim de subjectes que es toleraria que no superessin la prova? Se simbolitza per $f_{m\grave{a}x}$.
- 4) Quin és el percentatge mínim de subjectes que es toleraria que no superessin la prova? Se simbolitza per $f_{m\grave{i}n}$.

Com es veu, dues de les preguntes es basen en el nivell de coneixement que els avaluats han de presentar ($k_{m\grave{a}x}$ i $k_{m\grave{i}n}$). Aquestes dues qüestions es mesuren a partir del percentatge d'ítems que cal contestar correctament. Les altres dues preguntes es basen en el percentatge de no competents que es toleraria donada una determinada prova d'avaluació ($f_{m\grave{a}x}$ i $f_{m\grave{i}n}$). A partir d'aquests quatre punts i la distribució empírica de la prova es troba el punt de tall (x) òptim per a aquest mètode de compromís.

En la figura 3 es representa gràficament com es pot trobar aquest punt de tall. D'una banda, en l'eix d'abscisses es representa el percentatge d'ítems correctes de la prova, i en l'eix d'ordenades el percentatge de persones que no superen la prova. En aquesta gràfica també es representa la distribució empírica de la prova, la qual mostra com, a mesura que augmenta el nombre d'ítems que cal contestar correctament (eix d'abscisses), augmenta el percentatge de persones que no superen la prova. Els valors k i f es representen amb dos punts en l'eix de coordenades i s'uneixen amb una línia recta que creua la distribució empírica de la prova (en realitat s'espera que en la majoria dels casos travessarà la distribució empírica). El punt on es creua la distribució empírica i la recta k - f indicarà el percentatge d'ítems correctes que caldrà exigir a la prova (el punt de tall) i el percentatge corresponent de subjectes que no superen la prova si s'utilitza aquest punt de tall. Aquest punt de tall se simbolitza en la figura per x .

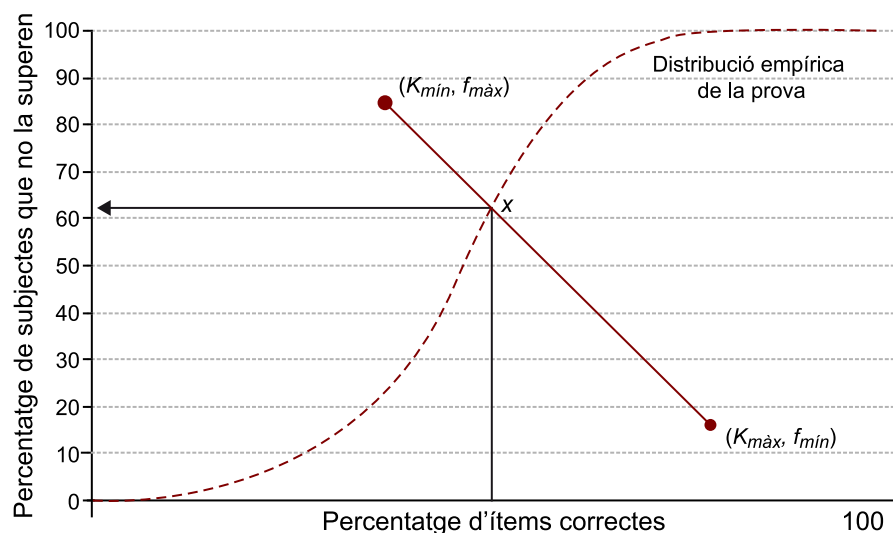


Figura 3. Representació gràfica del mètode de Hofstee

8.3.2. Mètode de Beuk

El mètode proposat per Beuk (1984) va sorgir com una simplificació del mètode de Hofstee (1983). Com l'anterior, pressuposa que els avaluadors tenen una idea més o menys clara sobre quin és el punt de tall que seria necessari aplicar a la prova i quin hauria de ser el percentatge de persones que haurien de superar-la. En aquest cas, es demana als avaluadors que responguin a les dues preguntes següents:

- 1) Quin hauria de ser el percentatge mínim d'ítems que caldria contestar correctament per a superar la prova? Aquest valor se simbolitza amb x .
- 2) Quin és el percentatge de persones que s'espera que superin la prova? Aquest valor se simbolitza amb y .

A partir del càlcul de la mitjana de les valoracions dels avaluadors d'aquestes dues preguntes ($\bar{x}$ i $\bar{y}$) i, d'altra banda, la distribució empírica obtinguda de l'aplicació de la prova, s'obté el punt de tall. L'exemple que es presenta a continuació i la figura 4 mostren com es pot obtenir el punt de tall a partir d'aquest mètode.

Suposem que com a mitjana els avaluadors han determinat que caldria que un 60% dels ítems de la prova es contestessin correctament ($\bar{x}$) i que un 70% de les persones que fessin la prova caldria que la superessin ($\bar{y}$).

En la figura 4 es representen aquests dos punts, i la seva intersecció se simbolitza amb la lletra A. Es representa la distribució empírica de la prova (línia de punts), que en aquest cas serà decreixent, ja que el nombre de persones que superaran la prova (eix d'ordenades) disminuirà a mesura que el percentatge d'ítems que cal contestar de manera correcta (eix d'abscisses) augmenti.

El pas següent és calcular les desviacions típiques en les dues preguntes formulades als avaluadors. El quocient entre les dues desviacions típiques (S_x/S_y) serà el pendent de la recta que travessarà la distribució empírica de la prova. El punt en què la recta travessa la distribució empírica se simbolitza amb el punt B. Finalment, com es pot veure en la figura 4, la projecció d'aquest punt a l'eix d'abscisses proporcionarà el punt de tall de la prova

(percentatge d'ítems correctes), i la seva projecció sobre l'eix d'ordenades proporcionarà el percentatge de subjectes que superaran la prova.

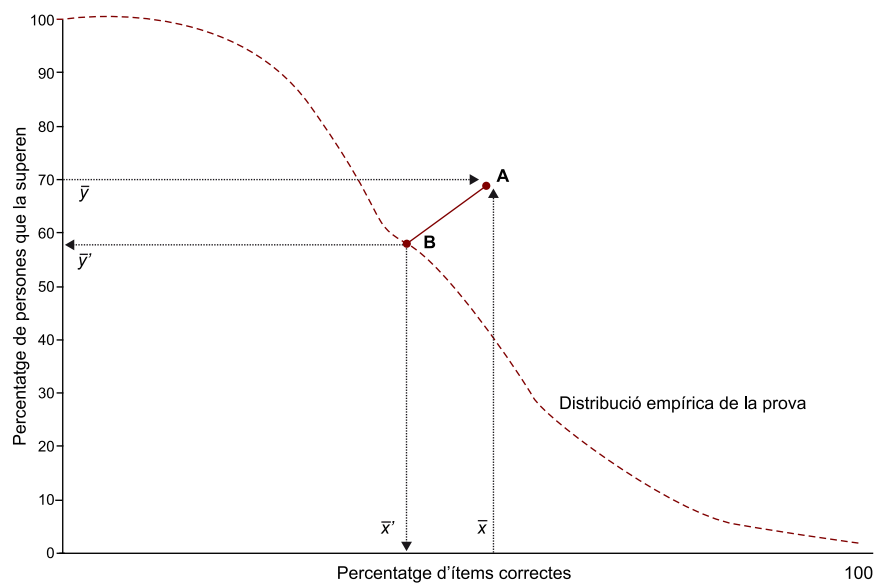


Figura 4. Representació gràfica del mètode de Beuk

Bibliografia

Angoff, W. H. (1971). Scales norms and equivalent scores. A R. L. Thorndike (Ed.). *Educational measurement* (2a. ed.). Washington: American Council on Education.

Angoff, W. H. (1984). *Scales norms and equivalent scores*. Princeton, NJ: Educational Testing Service.

Barbero, M. I., Vila, E., i Suárez, L. C. (2003). *Psicometría*. Madrid: UNED.

Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education*, 45(2), 4-9.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced test. *Review of Educational Research*, 56 (1), 137-172.

Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement*, 21 (2), 147-152.

Brennan, R. L. i Wan, L. (2004). *A bootstrap procedure for estimating decision consistency for single-administration complex assessments*. (CASMA Research Report, 7). Center for Advanced Studies in Measurement and Assessment / The University of Iowa.

Brennan, R. L. i Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement*, 14(3), 277-289.

Breyer, F. J., i Lewis, C. (1994). *Pass-fail reliability for tests with cut scores: A simplified method*. (ETS Research Report, 94-39). Princeton, NJ: Educational Testing Service.

Cizek, G. J. i Bunch, M. B. (2007). *Standard setting. A guide to establishing and evaluating performance standards on tests*. Thousand Oaks: SAGE Publications, Inc.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1), 37-46.

Cohen, R. J. i Swerdlik, M. E. (2009). *Psychological testing and assessment: An introduction to tests and measurement* (7a. ed.). Boston: McGraw-Hill.

Crocker L. i Algina, J. (1986). *Introduction to classical modern test theory*. Nova York: Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient alfa and the internal structures of tests. *Psychometrika*, 16(3), 297-334.

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika*, 30 (3), 357-370.

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder -Richardson coefficient twenty is the same for two tests. *Psychometrika*, 34 (3), 363-373.

Feldt, L. S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika*, 45 (1), 99-105.

Flanagan, J. C. (1937). A note on calculating the standard error of measurement and reliability coefficients with the test scoring machine. *Journal of Applied Psychology*, 23(4), 529.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10(4), 255-282.

Hambleton, R. K. i Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced test. *Journal of Educational Measurement*, 10 (3), 159-170.

Hambleton, R. K. i Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, 8 (1), 41-55.

Hofstee, W. K. (1983). The case for compromise in educational selection and grading. A S. B. Anderson, S. B. i Helmick, J. S. (Eds.). *On educational testing*. San Francisco: Jossey-Bass.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, 13(4), 253-264.

- Jaeger, R. M. (1989). Certification of student competence. A R. L. Linn (Ed.), *Educational measurement* (3a. ed.). Nova York: Macmillan.
- Kaplan, R. M. i Saccuzo, D. P. (2009). *Psychological testing. Principles, applications and issues* (7a. ed.). Belmont, CA: Wadsworth.
- Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika*, 28 (3), 221-238.
- Kuder, G. F. i Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2 (3), 151-160.
- Landis, J. R. i Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33 (1), 159-74.
- Lee, W. (2005). Classification consistency under the compound multinomial model. (CASMA Research Report, 13). Center for Advanced Studies in Measurement and Assessment / The University of Iowa.
- Lee, W., Brennan, R. L., i Wan, L. (2009). Classification consistency and accuracy for complex assessments under the compound multinomial model. *Applied Psychological Measurement*, 33 (5), 374-390.
- Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45 (1), 255-268.
- Livingston, S. A. (1972). Criterion-referenced applications of classical test theory. *Journal of Educational Measurement*, 9 (1), 13-26.
- Livingston, S. A. i Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32 (2), 179-197.
- Martínez-Arias, R. (2010). La evaluación del desempeño. *Papeles del Psicólogo*, 31 (1), 85-96.
- Muñiz, J. (2003). *Teoría clásica de los tests*. Madrid: Pirámide.
- Murphy, K. R. i Davidshofer, C. O. (2005). *Psychological testing. Principles and applications* (6a. ed.). Upper Saddle River, NJ: Prentice Hall.
- Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 14 (1), 3-19.
- Nunnally, J. C. (1978). *Psychometric theory* (2a. ed.). New York: McGraw-Hill.
- Plake, B. S. i Hambleton, R. K. (2001). The analytic judgment method for setting standards on complex performance assessments. A G. J. Cizek (Ed.). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum.
- Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*, 9, 99-103.
- Shrock, S. A. i Coscarelli, W. C. (2007). *Criterion-referenced test development: Technical and legal guidelines for corporate training and certification* (3a. ed.). San Francisco, CA: John Wiley and Sons.
- Sireci, S. G., Hambleton, R. K., i Pitoniak, M. J. (2004). Setting passing scores on licensure examinations using direct consensus. *CLEAR Exam Review*, 15 (1), 21-25.
- Subkoviak, M. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, 13 (4), 265-275.
- Woodruff, D. J. i Feldt, L. S. (1986). Test for equality of several alpha coefficients when their sample estimates are dependent. *Psychometrika*, 51 (3), 393-413.
- Zieky, M. J. i Livingston, S. A. (1977). *Manual for setting standards on the Basic Skills Assessment Tests*. Princeton, NJ: Educational Testing Service.

Validesa

PID_00216192

Luis Manuel Lozano
Jaume Turbany

Temps mínim de dedicació recomanat: 4 hores



**Luis Manuel Lozano**

Professor contractat doctor de l'Àrea de Metodologia de les Ciències del Comportament de la Universitat de Granada, membre del CIMCYC i consultor dels Estudis de Psicologia de la Universitat Oberta de Catalunya (UOC). Doctor en Psicologia per la Universitat d'Oviedo. Desenvolupa la docència en l'àrea de metodologia, i concretament imparteix les assignatures de Descripció i exploració de dades en psicologia i Tècniques d'anàlisi en recerques psicològiques. Els seus camps de recerca estan relacionats amb l'anàlisi freqüentista i bayesià de dades, l'elaboració de qüestionaris i models de teoria de la resposta als ítems.

**Jaume Turbany**

Professor titular d'universitat del Departament de Metodologia de les Ciències del Comportament de la Facultat de Psicologia de la Universitat de Barcelona (UB) i consultor dels Estudis de Psicologia i Ciències de l'Educació de la Universitat Oberta de Catalunya (UOC). Doctor en Psicologia per la UB. La seva tasca docent s'ha desenvolupat en l'àmbit de la metodologia i l'estadística aplicades a la psicologia i a la informació i documentació. Actualment la seva recerca s'orienta, entre altres camps, als estudis bibliomètrics en diferents àrees de coneixement, i a la recerca en innovació docent.

Índex

Introducció.....	5
1. Què és la validesa.....	7
1.1. Definició	7
1.2. Importància de la validesa	10
2. Evidència de validesa basada en el contingut.....	11
2.1. Concepte	11
2.2. Procediment	12
2.3. Contingut esbiaixat	13
3. Evidència de validesa basada en el procés de resposta.....	14
3.1. Concepte	14
3.2. Procediment	16
4. Evidència de validesa basada en l'estructura interna.....	17
4.1. Concepte	17
4.2. Procediments	18
4.2.1. Unidimensionalitat	18
4.2.2. Multidimensionalitat	22
5. Evidència de validesa basada en la relació amb altres variables.....	28
5.1. Concepte	28
5.2. Evidència de decisió (sensibilitat i especificitat)	29
5.3. Evidències convergents i/o discriminants	31
5.4. Evidències basades en les relacions test-criteri	33
5.4.1. Validesa concurrent o simultània	33
5.4.2. Validesa predictiva	35
5.4.3. Validesa retrospectiva	44
5.5. Generalització de la validesa	44
6. Evidència de validesa basada en les conseqüències de l'aplicació.....	46
7. Factors que afecten la validesa.....	47
7.1. Fórmules d'atenuació	47
7.1.1. Estimació del coeficient de validesa en el supòsit en què el test i el criteri tinguin una fiabilitat perfecta	47
7.1.2. Estimació del coeficient de validesa en el supòsit en què el test tingui una fiabilitat perfecta	48

7.1.3.	Estimació del coeficient de validesa en el supòsit en què el criteri tingui una fiabilitat perfecta	48
7.1.4.	Estimació del coeficient de validesa en el supòsit en què s'hagi millorat tant la fiabilitat del test com la del criteri	49
7.1.5.	Estimació del coeficient de validesa en el supòsit en què s'hagi millorat la fiabilitat del test	49
7.1.6.	Estimació del coeficient de validesa en el supòsit en què s'hagi millorat la fiabilitat del criteri	49
7.1.7.	Valor màxim que pot assolir el coeficient de correlació entre test i criteri	50
7.2.	Efecte de la longitud del test sobre el coeficient de correlació test-criteri	50
7.3.	Efecte de la variabilitat de la mostra en la correlació test-criteri	51
Bibliografia		53

Introducció

Quan un psicòleg decideix aplicar un qüestionari és per a aconseguir un objectiu determinat. Per a això s'ha d'assegurar que el qüestionari que usará té unes propietats psicomètriques adequades. Entre aquestes, cal destacar la propietat a la qual es fa referència en aquest mòdul: la validesa.

La validesa és un dels aspectes més importants, potser el que més, tant en l'elaboració com en l'avaluació de qüestionaris psicològics. Al cap i a la fi es tracta de comprovar que la utilització del test és correcta i que els objectius que vol assolir el psicòleg que l'utilitza són factibles.

En l'apartat “Què és la validesa?” es fa un breu recorregut històric sobre aquest concepte. Com es pot observar, és un concepte que ha estat evolucionant (i encara ho fa) fins que ha arribat a la idea que actualment està en vigor. Aquest concepte es defineix oficialment en els Standards publicats el 1999 conjuntament per l'American Educational Research Methods (AERA), l'American Psychological Association (APA) i el National Council on Measurement in Education (NCME). Aquestes entitats defensen que es poden agrupar els indicis de validesa d'un test en cinc apartats: evidència basada en la validesa de contingut, basada en el procés de resposta, basada en l'estructura interna del qüestionari, basada en la relació amb altres variables i basada en les conseqüències de l'avaluació.

En els apartats següents es veu cadascun dels indicis de validesa prèviament esmentats i es busquen estratègies per a poder obtenir aquests indicis (a excepció de l'apartat de les conseqüències de l'avaluació, en el qual només es tracta de les conseqüències que es poden esperar de l'aplicació d'un qüestionari).

En l'últim apartat, “Factors que afecten la validesa”, es veuen diferents aspectes que afecten algunes de les tècniques exposades amb anterioritat per a determinar els diferents indicis de validesa.

1. Què és la validesa

1.1. Definició

Per a comprendre el concepte de *validesa* cal fer un petit estudi de l'evolució històrica que ha experimentat aquest concepte.

La utilització de qüestionaris es va veure impulsada per la Primera i la Segona Guerra Mundial. En aquells moments es va tenir la necessitat d'incorporar la població civil a l'exèrcit, destinant cada persona al lloc més adequat per a ella. Després d'emplenar els qüestionaris es comprovava al camp d'entrenament si els subjectes rendien satisfactòriament o no al lloc on havien estat destinats. Atès que en primer lloc es feia el mesurament i posteriorment s'avaluava l'èxit, es parlava de validesa predictiva. És a dir, un test té **validesa predictiva** si serveix per a predir el comportament en un constructe que serà avaluat després d'aplicar el qüestionari.

Posteriorment, es va tractar d'avaluar la relació entre les característiques de les persones que feien un treball i el seu èxit en aquest. D'aquesta manera, es tractava de saber quines característiques podrien predir l'èxit laboral i buscar-les quan es feia una selecció de personal. Atès que l'estudi s'efectuava sobre persones que ja tenien el lloc i se'n valorava l'execució, es parlava de validesa concurrent, ja que tots dos mesuraments es feien alhora. És a dir, un test té **validesa concurrent** si serveix per a predir el comportament en un constructe que és avaluat simultàniament a l'aplicació del qüestionari.

Com es pot observar, inicialment els tests s'empraven exclusivament per a predir. Així doncs, en un començament, es considerava que un test era vàlid si servia per a predir alguna variable d'interès denominada **criteri** (Guilford, 1946).

Per tant, es conceptualitza la validesa com a correlació entre el qüestionari i el criteri d'interès (tant si aquest és avaluat amb posterioritat com simultàniament a l'aplicació del qüestionari). Així doncs, es considera que un test és vàlid per a avaluar qualsevol aspecte amb el qual es correlacioni (Bingham, 1937; Guilford, 1946; entre d'altres).

Un dels problemes de la conceptualització de la validesa com a correlació és el fet que cal trobar una mesura del criteri adequada, és a dir, es necessiten dades del criteri que s'hagin obtingut d'una manera fiable i vàlida. Per tant, si ja es disposa d'una mesura vàlida del criteri per a què es necessita aplicar un qüestionari?

Un altre problema d'aquesta conceptualització és que deixava fora un gran nombre de tests educatius. En aquests no es tracta de predir la conducta, es tracta de comprovar quant s'ha après després d'un període de formació. En aquests qüestionaris la puntuació obtinguda és un indicador del que el test vol avaluar (coneixement en matemàtiques, en anglès, etc.) i no un predictor de criteris diferents del test. Des d'aquesta perspectiva, la validesa fa referència al fet que els ítems que componen el qüestionari siguin representatius d'allò que es vol avaluar. Aquest concepte es va denominar **validesa de contingut** (Anastasi, 1954).

D'altra banda, al llarg dels anys trenta hi ha un auge de les teories que tracten de conèixer l'estructura factorial de la intel·ligència. Amb aquestes teories es comença a conceptualitzar un test com a vàlid quan representa de manera fidedigna el constructe psicològic que vol mesurar, com també les relacions esperades entre els diferents constructes. D'aquesta manera neix la validesa de constructe (Cronbach i Meehl, 1955). Les tècniques estadístiques emprades per a poder comprovar aquesta validesa són, tradicionalment, l'anàlisi factorial exploratòria i les matrius multitret-multimètode (Campbell i Fiske, 1959), i més recentment l'anàlisi factorial confirmatòria. Per exemple, si s'empra un test que avalua la tríada cognitiva des del model cognitiu de depressió de Beck (Beck, Rush, Shawn i Emery, 1979) (pensaments sobre un mateix, pensaments sobre el món i pensaments sobre el futur), el qüestionari tindrà **validesa de constructe** si avalua les tres dimensions i aquestes tenen les relacions que s'esperen, per exemple, amb ansietat.

Fins als anys vuitanta es podia parlar de validesa predictiva, validesa concurrent, validesa de contingut i validesa de constructe d'un qüestionari. Si bé les dues primeres en els estàndards dels tests i manuals educatius i psicològics publicats per l'APA, AERA i NCME el 1966 i el 1974 s'englobaven com a **validesa de criteri**.

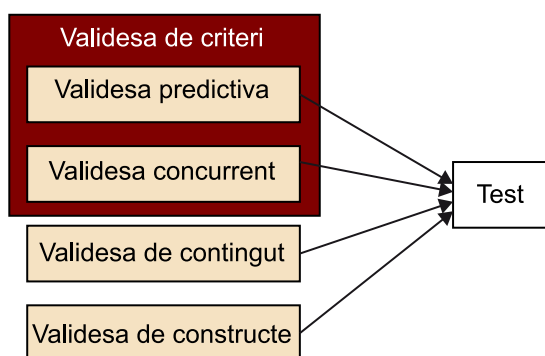


Figura 1

Posteriorment, Cronbach (1971) va puntualitzar que en un test que vol mesurar un tret de personalitat no hi ha només un criteri rellevant per a predir ni un contingut per a mostrejar (validesa predictiva i de contingut respectivament). Per contra, es disposa d'una teoria sobre el tret i sobre les seves relacions amb altres constructes i variables (validesa de constructe). Si se suposa que

la puntuació del test és una manifestació vàlida de l'atribut, es pot contrastar l'assumpció analitzant les seves relacions amb altres variables. Per tant, va començar a haver-hi una tendència que considerava la validesa com una cosa unitària, essent la validesa de constructe la científicament més admissible i estant la validesa de criteri i de contingut incloses en aquesta (Messick, 1989). Així doncs, s'imposa la concepció que la validació de constructe constitueix un marc integral per a obtenir proves de la validesa incloent-hi les procedents de la validació de criteri i de contingut. De fet, es deixa de parlar de les diferents categories de validesa i es comença a parlar de diferents evidències implicades en els tres tipus tradicionals de validesa (criteri, contingut i constructe).

Atès que tant l'estudi de l'estructura del constructe com les relacions d'aquest amb altres constructes es passa a considerar la forma principal de validesa, aquest procés es pot concebre com un cas particular de la contrastació de les teories científiques mitjançant el mètode hipoteticodeductiu (Prieto i Delgado, 2010).

Fixeu-vos que en aquests moments, a mitjan anys vuitanta, hi ha un canvi molt rellevant. Mentre que al començament es conceptualitza la validesa com una propietat inherent al test, es passa a concebre que el que realment es valida no és el test en si mateix, sinó les inferències que es fan a partir d'aquest. Per això, el responsable d'assegurar la validesa ja no és només el constructor del test, sinó que també ho és l'usuari que empra aquest qüestionari per a una finalitat determinada. Sovint, els problemes d'un qüestionari sobre la validesa es deuen no al disseny del qüestionari, sinó a l'ús que se'n fa.

Actualment, en l'última edició fins ara dels *Standards for educational and psychological testing* (AERA, APA, NCME, 1999), molt influenciats pel capítol escrit per Messick (1989) i el llibre de Shepard, Camilli, Linn i Bohrnstedt (1993), es defensa que la validesa fa referència al **grau en què l'evidència empírica i la teoria donen suport a la interpretació de les puntuacions dels tests relacionada amb un ús específic**. Com es pot apreciar, la validesa es concep com un concepte unitari. Per a comprovar la validesa s'ha d'atendre a cinc evidències de validesa:

- **El contingut de test:** els ítems que constitueixen el test són rellevants i representatius del constructe psicològic que es vol mesurar.
- **El procés de resposta:** el procés que segueixen les persones en contestar el test permet extreure respostes indicadores del que es vol avaluar.
- **L'estructura interna:** les relacions dels ítems entre ells són congruents amb el model teòric emprat a l'hora de definir el constructe que es vol avaluar.

- **La relació amb altres variables:** les relacions que s'estableixen entre el constructe que s'avalua i altres constructes són les esperades segons el marc teòric en el qual s'ha definit el constructe a avaluar.
- **Les conseqüències de l'aplicació del qüestionari:** les conseqüències tant positives com negatives que ha d'emprar un test són les previstes.

Com a breu resum del que s'ha exposat més amunt, es presenta la taula següent, en la qual es pot apreciar l'evolució del concepte en els diferents estàndards publicats per l'APA.

Taula 1

Edició	Validesa
1954	Constructe, concurrent, predictiva, contingut
1966	Criteri, constructe, contingut
1974	Criteri, constructe, contingut
1985	Unitària (però mantenen criteri, constructe i contingut)
1999	Unitària: 5 fonts d'evidència

1.2. Importància de la validesa

El concepte de *validesa* és central en la psicometria. Tal com s'ha comentat anteriorment, per a comprovar la validesa s'han d'acumular evidències que proporcionin una base científica per a interpretar les puntuacions d'un qüestionari de manera adequada. Per això el que realment es valida no és el qüestionari en si mateix, són les interpretacions que es fan a partir d'aquest. Per tant, no es pot defensar que un test sigui vàlid o per contra manqui de validesa. Un test pot ser adequat per a un propòsit però no per a un altre.

Si s'aplica un qüestionari amb el qual es vol mesurar l'autoestima, les respostes es poden emprar amb diferents finalitats (conèixer el nivell d'autoestima d'una persona per a saber si és un problema que s'ha de tractar en teràpia, en selecció de personal, com a recerca sobre el propi constructe, etc.). Per a poder usar el qüestionari amb una finalitat determinada, s'han d'acumular evidències que indiquin que l'ús és correcte (evidències de validesa). En cas contrari, s'estaria fent un mal ús dels tests, eines principals en el treball psicològic, i les conclusions que se'n traguessin no serien correctes. En l'exemple anterior no se sabia si és un aspecte sobre el qual s'ha d'intervenir terapèuticament, no se sabia si la persona seleccionada realment té l'autoestima que es desitja o no se sabia si realment s'està mesurant autoestima.

Per a poder fer el treball correctament com a psicòlegs, s'ha de saber si les conclusions que es treuen a partir dels tests emprats són adequades, ja que, en cas contrari, es corre el risc de no saber exactament què s'està avaluant o si aquest mesurament realment és útil per al propòsit del psicòleg.

2. Evidència de validesa basada en el contingut

2.1. Concepte

Moltes de les inferències i assumpcions que es deriven de la interpretació de les puntuacions en un test es poden avaluar més fàcilment si s'examinen els procediments que s'han emprat per a generar les puntuacions. Per exemple, si es vol inferir alguna cosa a partir de les puntuacions en un test sobre determinada conducta o constructe psicològic, se suposa que els ítems que componen el qüestionari són tan **rellevants** (que la informació que es demana està directament relacionada amb el que es vol mesurar) com **representatius** (les qüestions que es presenten han de ser una mostra adequada de tot el que es vol mesurar) d'aquesta (Kane, 2006).

L'evidència de la validesa de contingut fa referència a la relació que hi ha entre els ítems que componen el test i el que es vol avaluar amb aquest, parant esment tant a la rellevància com a la representativitat dels ítems. Aquest tipus d'evidència es recull principalment en el moment de l'elaboració del test.

Suposem que es vol elaborar un test per a avaluar la personalitat. En aquest cas, es decideix treballar dins del marc teòric dels cinc factors de la personalitat (extraversió, obertura, responsabilitat, amabilitat i neuroticisme). Atès que es tracta d'un test que s'emprarà en una selecció de personal concreta, només interessen les dimensions de responsabilitat (a), amabilitat (b) i neuroticisme (c). En aquest exemple, el constructe és la personalitat que està composta per les cinc dimensions. Les dues primeres, per als interessos del test que es planteja, són informació irrellevant. Les altres tres són el domini que interessa avaluar. A partir d'aquest domini es construeixen ítems destinats a avaluar la responsabilitat (a'), l'amabilitat (b') i el neuroticisme (c'). Aquests ítems han de tenir relació amb el factor que volem mesurar, és a dir, els ítems que avaluen responsabilitat estan relacionats amb la definició que existeix en la comunitat científica sobre aquest factor (rellevància). Però, al seu torn, els ítems han de preguntar per la totalitat del domini que s'ha d'avaluar (representativitat).

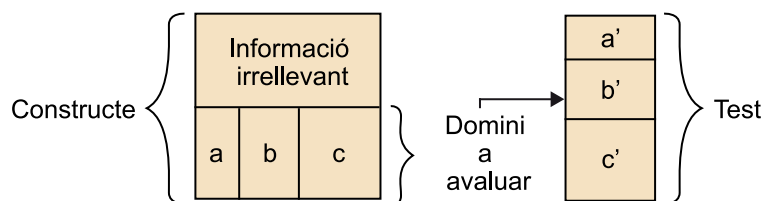


Figura 2

En les proves educatives, les evidències de validesa basada en el contingut són fonamentals. Si no es comprova que el test és consistent amb els objectius curriculars que es persegueixen (rellevància), és a dir que està lliure de material irrellevant i que el que hi ha representa adequadament el domini que es vol avaluar (representativitat), la utilitat del test es veurà seriosament afectada i, per tant, les conclusions que s'obtinguin a partir d'aquest seran errònies. En aquestes situacions, atès que el domini que es vol avaluar està clarament defi-

nit, se sol recomanar emprar els diferents mètodes estadístics de mostreig per a obtenir una mostra representativa dels continguts que han de constituir el test (Muñiz, 2003).

El problema sorgeix quan no es disposa del domini tan clarament definit.

Per exemple, si es vol elaborar un test que avaluï la intel·ligència, el primer que s'ha de preguntar el constructor del qüestionari és què és la conducta intel·ligent. En aquest cas, atès que no hi ha un domini perfectament definit, s'han de buscar altres estratègies per a obtenir l'indicador de la validesa de contingut.

2.2. Procediment

En aquest apartat es presentarà el procediment més habitual en la valoració de l'evidència basada en el contingut, si bé hi ha altres mètodes menys emprats. En podeu trobar una recopilació a Sireci (1998).

Si es vol desenvolupar un test, el primer que s'ha de fer és definir de manera operativa el domini que es vol avaluar. Després d'elaborar-ne una o acceptar-ne una que ja existeix, s'ha d'elaborar una **taula d'especificacions**. Es tracta de fer una descripció detallada del test, determinar la proporció o el nombre d'ítems que avaluaran cada contingut o habilitat del domini que es vol avaluar i determinar el format dels ítems i de les respostes (AERA, APA, NCME, 1999) (usualment en aquest pas també es determinen les propietats psicomètriques que es vol que tingui la prova).

Després de crear els ítems s'ha d'acudir a un grup d'experts en la matèria que faran de jutges. Per a evitar qualsevol biaix, aquests jutges no han d'estar implicats en l'elaboració del qüestionari, sinó que han d'analitzar cadascun dels ítems valorant en quina mesura són **representatius i rellevants** per a avaluar el domini d'interès, prenent com a definició d'aquest l'aportada pels autors del test.

Es pot defensar que hi ha, per tant, tres aspectes ben diferenciats que s'han de tenir en compte a l'hora de comprovar les evidències de la validesa de contingut: la definició del domini, la representació dels ítems que avaluen el domini i la seva rellevància (Sireci, 1998).

És recomanable que cada jutge valori els ítems per separat per a evitar, d'aquesta manera, possibles biaixos a l'hora de respondre. Una vegada es tenen les valoracions de tots els experts, s'han de buscar els ítems en els quals hi hagi concordança i seleccionar-los per a formar part del qüestionari.

Per exemple, si 8 dels 10 jutges determinen que un ítem destinat a mesurar depressió realment avalua allò que vol avaluar, aquest ítem tindrà un índex de congruència de 0,8. Se solen considerar adequats els ítems que tenen un índex de congruència igual o superior a 0,7 (Sireci, 1998).

No cal eliminar els ítems en els quals no hi hagi acord (que no aconseguixin un índex de congruència de 0,7). És recomanable que amb aquests ítems es faci un grup de discussió

amb els experts perquè comentin les diferències tractant d'arribar a un punt d'acord per a millorar aquests ítems.

Aquest és el procediment més habitual a l'hora de valorar els indicis de validesa de contingut, si bé no està lliure de crítiques. El problema principal que es planteja en la utilització d'experts és que aquests són altament competents en el contingut que s'avalua, per la qual cosa poden passar per alt un text amb un nivell que no és adequat per a la comprensió dels subjectes que s'han d'avaluar o que es pot interpretar malament amb facilitat. És a dir, encara que l'expert ens pot proporcionar informació molt rellevant, el que realment importa és com percep el test o l'ítem la persona que hi està responnent i com hi reacciona (Leighton, 2004).

2.3. Contingut esbiaixat

L'ús d'experts per a valorar tant la rellevància com la representativitat dels ítems té com a finalitat evitar que el qüestionari tingui continguts esbiaixats. Es diu que el contingut d'un test està esbiaixat si els ítems que el componen avaluen aspectes no rellevants per al domini (**biaix per falta de rellevància**) o si no representen de manera adequada tot el domini a avaluar (**biaix per falta de representativitat**). Com es pot comprovar, un test està esbiaixat si no cobreix adequadament el domini que vol mesurar o si inclou qüestions que no són necessàries per a valorar el domini correctament.

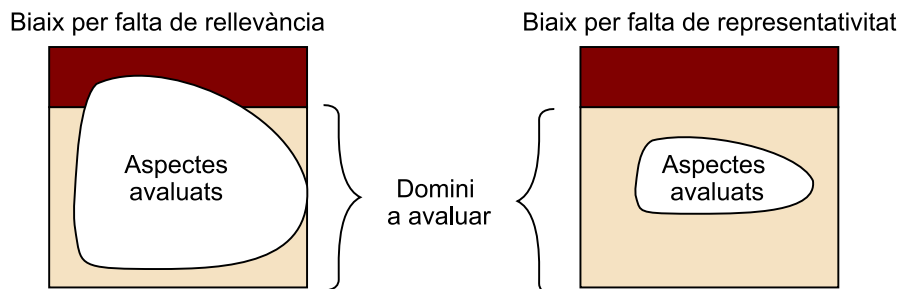


Figura 3

3. Evidència de validesa basada en el procés de resposta

3.1. Concepte

Aquest tipus d'evidència és un concepte que es va introduir com a nou en els *Standards* publicats el 1999, si bé ja havia estat esmentat per alguns especialistes en la mesura del psicològic com Messick (1989). Els *Standards* descriuen aquest tipus d'indicis com l'ajust entre el constructe avaluat i el procés de resposta que han dut a terme les persones que responen al test.

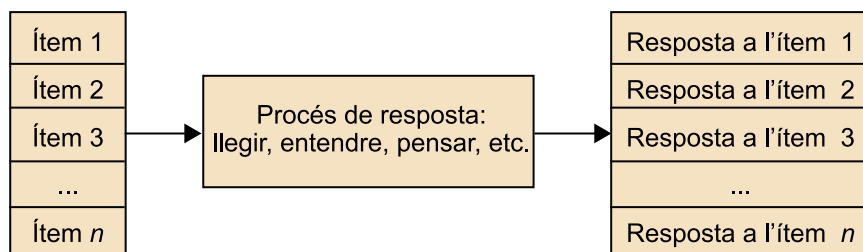


Figura 4

Per **procés de resposta** s'entenen totes les conductes que es necessiten per a poder contestar un ítem, com poden ser llegir les preguntes, comprendre-les, decidir la resposta que es vol donar i finalment respondre a l'ítem.

Un exemple sobre aquest índex de validesa es pot trobar en un examen de matemàtiques que es faci a nens que estan aprenent a sumar. Si l'enunciat de l'ítem és " $3 + 2 =$ ", probablement, si han adquirit el coneixement necessari, poden donar una resposta correcta. Aquest ítem també pot tenir un enunciat com "Calculeu el valor resultant de dur a terme una operació additiva entre els valors 3 i 2". Evidentment, un nen que estigui aprenent a sumar pot respondre al primer enunciat, però no al segon (no té el vocabulari necessari, la seva capacitat lectora no li permetrà comprendre la pregunta, etc.). Com a resultat del primer enunciat es conclourà que ja té adquirit cert nivell del domini avaluat, però amb el segon es conclourà que no. És a dir, atès que el segon enunciat manca de validesa de resposta, portarà els avaluadors a conclusions errònies sobre el nivell d'habilitat del nen a l'hora de fer sumes.

A l'hora de respondre a un test, s'han de combinar tant les característiques de les preguntes com les de les respostes que es poden donar i les de la persona que respon. Per això hi ha diferents factors que poden afectar la resposta:

- **Factors relacionats amb els ítems.** En aquest apartat s'han de tenir en compte diversos factors.
 - Contingut dels ítems. S'ha d'assegurar que el contingut és adequat a la població que es vol avaluar. Per exemple, no es pot usar un test per a avaluar depressió infantil si es va construir per a avaluar-la en adults.

Si es fa, es poden trobar preguntes del tipus “Ha notat canvis en el seu desig sexual?”, que evidentment són inadequades per a avaluar nens.

- Redacció dels ítems. El llenguatge emprat per a redactar l'ítem no ha de superar la capacitat comprensiva de les persones que respondran. Un exemple d'això es pot observar en l'exemple de la suma que hem exposat anteriorment.
- Validesa aparent de l'ítem. Quan s'avaluen coneixements és desitjable que les persones que responen al qüestionari pensin que és adequat. Si s'avalua el coneixement en psicometria mitjançant un test, s'espera que els alumnes que facin el test pensin que les preguntes són adequades per a mesurar el coneixement en psicometria. Això és el que es denomina *validesa aparent*. Per contra, en els tests de personalitat s'ha d'intentar que la persona que respon no sàpiga exactament el que s'avalua. D'aquesta manera s'intenta evitar que respongui el que més l'afavoreix o el que pensa que s'espera d'ell.
- **Factors relacionats amb la resposta als ítems**
 - El nombre d'alternatives que s'ofereixin com a resposta. Als tests d'actituds se sol respondre en un format tipus Likert. En aquestes escales es demana a les persones en quin grau estan d'acord amb l'afirmació que se'ls presenta i han de respondre en una escala en què, per exemple, 0 significa totalment en desacord i 5 totalment d'acord. El problema sorgeix quan s'empra una escala que supera la capacitat discriminativa de les persones. En estudiants universitaris, una escala de 0 a 10 és perfectament comprensible però, per contra, aquesta mateixa escala emprada en persones sense estudis pot ser excessiva. Un universitari comprèn perfectament la diferència entre un 4 i un 5, però aquesta diferència pot ser menys clara en una persona sense estudis, de manera que s'introduiria un error en l'avaluació.
 - Les instruccions a l'hora d'emplenar el qüestionari. A l'hora d'emplenar un qüestionari, les instruccions han de ser clares i comprensibles. S'han d'adaptar al grup que es vol avaluar perquè el criteri emprat a l'hora de respondre sigui clar.
- **Factors relacionats amb les persones.** En aquest apartat entrarien totes les característiques personals dels que respondran al qüestionari (capacitat lectora, capacitat intel·lectual, capacitat discriminativa, estat emocional, etc.). Cal fer esment especial de les situacions en què la persona està en un procés de selecció, ja que tractarà de donar una imatge distorsionada de si mateix, tractant d'adaptar-se al que pensa que el seleccionador busca.

3.2. Procediment

Encara que en els *Standards* de l'APA (1999) apareix aquest índex de validesa, amb prou feines aporten informació sobre com es pot determinar si un test té índexs de validesa basat en el procés de resposta. Dins de les alternatives que proposen hi ha tècniques com les següents:

- **L'entrevista.** Es pregunta a les persones que responen al test per les diferents estratègies emprades per a contestar cadascun dels ítems. El coneixement d'aquestes estratègies pot conduir fins i tot a l'enriquiment de la definició del constructe estudiat.
- **Tècniques de pensament en veu alta.** Es demana a les persones que emplenin el qüestionari dient en veu alta els diferents processos pels quals passen a mesura que han de contestar el test.
- **Entrevistes cognitives.** Estan dissenyades per a comprendre com les persones que responen a un test comprenen la pregunta, recuperen la informació rellevant per a respondre, avaluen la rellevància del recordat i responen a la pregunta. Emprant aquesta informació es poden identificar errors de resposta potencials i patrons d'interpretació de les preguntes. També poden aportar informació sobre els factors socioculturals que afecten la manera de respondre (Czaja i Blair, 1996).

4. Evidència de validesa basada en l'estructura interna

4.1. Concepte

Per a elaborar un test s'utilitzaran diversos ítems o preguntes. És possible que es consideri que tots els ítems són igual de rellevants per a mesurar la característica que s'estudia, en aquest cas obtindrem una puntuació total del test per la simple suma de les puntuacions que ha obtingut el subjecte en els diferents ítems.

La situació pot no ser tan senzilla quan suposem que no tots els ítems tenen la mateixa importància en la mesura del constructe, i per tant caldrà ponderar les puntuacions dels ítems, abans de procedir a la suma: en aquest cas parlarem de *puntuacions compostes*. En aquesta situació, l'estructura del test que haurem de determinar és unidimensional, ja que suposem que tots els ítems, encara que de diverses maneres, contribueixen a la mesura d'un únic aspecte de la variable criteri.

Un test també pot presentar una estructura interna multidimensional, això és, que les diferents preguntes no mesuren un sol aspecte sinó dues dimensions o més.

La tècnica estadística de l'anàlisi factorial ens servirà per a estudiar la contribució dels diferents ítems a un sol factor (estructura unidimensional) o a diversos factors (estructura multidimensional).

La tècnica de l'anàlisi factorial ens permetrà determinar k factors subjacents, a partir d'una sèrie p de puntuacions determinades pels ítems inicials del test. La idea és la cerca d'un model parsimoniós (simple) a partir d'un conjunt complex de dades.

A partir dels treballs de Spearman al principi del segle XX, i sobretot de Thurstone als anys quaranta del segle passat, l'anàlisi factorial s'evidencia com una bona eina en psicologia per a tractar d'identificar els factors que intervenen en la intel·ligència. Thurstone va proposar la utilització de l'anàlisi factorial per a explicar les correlacions que observava entre diferents ítems dels test d'intel·ligència. Així, l'ús d'aquesta tècnica li va permetre identificar i diferenciar les capacitats espacial, verbal i numèrica com a factors de la intel·ligència.

El problema d'aquesta tècnica rau en les dificultats del càlcul, sobretot a partir d'un nombre elevat d'ítems (variables). No obstant això, el desenvolupament i popularització actual dels programes estadístics ha permès la difusió d'aquesta i altres tècniques d'anàlisi de dades multivariables.

4.2. Procediments

El terme d'*anàlisi factorial* no designa un concepte unitari, sinó que reuneix diferents procediments que persegueixen la reducció inicial de múltiples variables en un nombre inferior de factors. En processos exploratoris, la tècnica més utilitzada és la de l'anàlisi en components principals, si bé hi ha altres formes d'extracció dels factors o components.

4.2.1. Unidimensionalitat

L'anàlisi en components principals parteix inicialment de la matriu de correlacions entre les diferents variables. Disposem de la matriu de correlacions obtinguda a partir de l'administració d'un test a una mostra de cinquanta-dos individus, i compost per vuit preguntes o ítems que intenten mesurar un únic constructe, en aquest cas l'autoestima dels subjectes.

Taula 2. *Correlation matrix*

	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6	Ítem 7	Ítem 8
Ítem 1	1,000	,447	,411	,444	,533	,337	,365	,442
Ítem 2	,447	1,000	,561	,662	,707	,528	,333	,522
Ítem 3	,411	,561	1,000	,665	,699	,462	,572	,540
Ítem 4	,444	,662	,665	1,000	,682	,518	,560	,564
Ítem 5	,533	,707	,699	,682	1,000	,467	,592	,488
Ítem 6	,337	,528	,462	,518	,467	1,000	,424	,418
Ítem 7	,365	,333	,572	,560	,592	,424	1,000	,422
Ítem 8	,442	,522	,540	,564	,488	,418	,422	1,000

En aquesta situació, la matriu de correlacions presenta una distribució de valors prou uniforme, i en aquest cas no es detecten agrupacions de variables amb correlacions altes entre elles i baixes amb les altres.

És possible descompondre la variància de cada variable (ítem) en tres fonts de variació: la variància factorial comuna, que comparteixen les variables en comú, la variància específica, o no compartida per altres variables, i la variància de l'error. La variància comuna, també denominada *comunalitat* (h^2), interessa que sigui prou alta una vegada hem seleccionat els factors rellevants.

En l'inici de l'anàlisi, la comunalitat de les variables és la unitat; després de l'anàlisi, com més a prop sigui d'1, més relació indicarà amb el factor o factors extrets.

Taula 3. Comunalitats

	Inicial	Extracció
Ítem 1	1,000	,410
Ítem 2	1,000	,628
Ítem 3	1,000	,670
Ítem 4	1,000	,722
Ítem 5	1,000	,743
Ítem 6	1,000	,455
Ítem 7	1,000	,489
Ítem 8	1,000	,518

Mètode d'extracció: anàlisi de components principals

La comunalitat d'un ítem j està representada per:

$$h_j^2 = a_j^2 + b_j^2 + \dots + k_j^2$$

En què a_j^2 , b_j^2 , ..., k_j^2 representen el quadrat dels coeficients de saturació de cada ítem amb cada factor A, B, ..., K, extrets, i el coeficient de saturació és la correlació de cada ítem amb els factors extrets.

En l'exemple les variables que presenten més comunalitat són els ítems 4 i 5.

Taula 4. Variància total explicada

Com- ponent	Autovalors inicials			Sumes de les saturacions al quadrat de l'extracció		
	Total	% de la variància	% acu- mulat	Total	% de la variància	% acumulat
1	4,636	57,947	57,947	4,636	57,947	57,947
2	,718	8,969	66,916			
3	,681	8,513	75,429			
4	,572	7,155	82,584			
5	,558	6,969	89,553			
6	,344	4,303	93,856			
7	,304	3,803	97,659			

Mètode d'extracció: anàlisi de components principals

Com- ponent	Autovalors inicials			Sumes de les saturacions al quadrat de l'extracció		
	Total	% de la variància	% acu- mulat	Total	% de la variància	% acumulat
8	,187	2,341	100,000			

Mètode d'extracció: anàlisi de components principals

A partir de p variables l'anàlisi factorial extreu el mateix nombre de factors, no relacionats entre ells. Cadascun dels factors es defineix com a combinació lineal de les p variables originals. Aquests p factors s'ordenen per ordre d'importància. En efecte, el primer component o factor és el millor resum de les relacions lineals que presenten les dades. El segon factor és la segona millor combinació de les variables, amb la condició que sigui ortogonal (sense relació) amb el primer, i així successivament amb la resta dels p factors o components.

Un criteri molt estès a l'hora d'extreure els components és el del valor propi o autovalor superior a 1, que és el més utilitzat. Un altre seria retenir els factors necessaris fins a aconseguir un percentatge adequat de variabilitat explicada pels components.

El valor propi o autovalor (λ) es defineix com la suma dels quadrats de les saturacions o correlacions de cada ítem amb el component en qüestió. Per tant, representa una mesura de la variabilitat explicada en les variables per part del component o factor.

En la taula de variància total explicada en què es desglossen els diferents components, veiem que solament hi ha un component amb valor propi (4,63) superior a 1. Per tant, confirmaria una estructura unidimensional del test, ja que totes les preguntes conflueixen en un sol component, identificable com el constructe subjacent, que es vol mesurar. En el nostre exemple, les diferents preguntes de l'escala elaborada contribuirien a la mesura d'un únic constructe psicològic que identificaríem amb l'autoestima.

Taula 5. Matriu de components^a

	Component
	1
Ítem 1	,640
Ítem 2	,793
Ítem 3	,819
Ítem 4	,850
Ítem 5	,862

Mètode d'extracció: anàlisi de components principals. a. 1 components extrets

	Component
	1
Ítem 6	,675
Ítem 7	,699
Ítem 8	,720

Mètode d'extracció: anàlisi de components principals. a. 1 components extrets

La matriu de components indica les correlacions entre cada ítem amb el component. Són les que abans hem denominat *saturacions* (*factor loadings*). En l'exemple veiem que els valors són alts, i a més amb poca fluctuació, la qual cosa indicaria que tots els ítems tenen una importància similar en la mesura del constructe.

Amb tots els indicadors esmentats podem comprovar que la comunalitat de cada ítem, en haver seleccionat un sol factor, simplement és el quadrat de la saturació entre ítem i factor.

Així, per a l'ítem 1, la comunalitat final és $h_1^2 = (0,64)^2 = 0,41$. Un 41% de la variabilitat del primer ítem és explicada pel component.

El valor propi del component 1 s'obté de la suma dels quadrats de les saturacions de cada ítem amb el factor.

Així en el primer component, $\lambda_1 = (0,64)^2 + (0,793)^2 + \dots + (0,72)^2 = 4,63$.

Com que tenim vuit ítems, el màxim serien vuit components. Si fem el quocient $4,63/8 = 0,5795$. Un 57,95% de la variabilitat total és explicada pel primer component.

Una vegada hem extret els components, disposarem de la matriu de puntuacions factorials. Aquesta matriu ens proporciona les ponderacions de cada variable per al càlcul de la puntuació de cada subjecte en els factors extrets. Les puntuacions factorials (*factor scores*) per a les dades individuals es calculen a partir de la matriu de coeficients de puntuacions factorials

$$F_i = a_1 \cdot Z_1 + a_2 \cdot Z_2 + \dots + a_p \cdot Z_p$$

Amb les dades de l'exemple, la matriu de ponderacions:

Lectura de la fórmula

a_i : coeficients de ponderació de cada variable per a cada factor

Z_i : puntuacions tipificades dels valors de cada variable obtinguts per cada individu.

Taula 6. Matriu de coeficients per al càlcul de les puntuacions en els components

	Component
	1
Ítem 1	,138
Ítem 2	,171
Ítem 3	,177
Ítem 4	,183
Ítem 5	,186
Ítem 6	,146
Ítem 7	,151
Ítem 8	,155

Mètode d'extracció: anàlisi de components principals

Per a cada subjecte és possible calcular una puntuació factorial de l'índex d'autoestima:

$$F_1 = 0,138 \cdot Z_{ítem1} + 0,171 \cdot Z_{ítem2} + \dots + 0,155 \cdot Z_{ítem8}$$

De totes maneres, en aquest exemple es veu que les ponderacions de tots els ítems són molt similars i que la contribució de totes les preguntes en la mesura del constructe d'interès és molt similar, i per tant seria adequat optar per una puntuació simple sumant les puntuacions obtingudes en cada ítem, contra una puntuació composta ponderant els valors de cada ítem.

4.2.2. Multidimensionalitat

Sovint, encara que d'entrada intentem elaborar una escala per a mesurar un sol constructe psicològic, és possible que després de la primera administració del test, en la prova pilot, observem que en realitat els ítems s'agrupen entre ells i afecten diferents constructes subjacents.

Presentem un altre exemple en el qual s'ha elaborat un qüestionari amb la intenció de mesurar les actituds sobre idees religioses en una mostra de 870 subjectes. Els ítems s'han identificat amb el concepte principal que implicava la pregunta. La matriu de correlacions de Pearson entre els ítems es presenta a continuació:

Taula 7. Correlation matrix

	Sentit vida	Religió	Obediència	Més enllà	Exper.	Inseguretat	Influència	Independ.
Sentit de la vida	1,000	,295	,220	,253	,226	,134	,178	,027
Religió	,295	1,000	,440	,507	,243	,099	,241	,103

	Sentit vida	Religió	Obediència	Més enllà	Exper.	Inseguretat	Influència	Independ.
Obediència	,220	,440	1,000	,339	,292	,063	,336	,054
Més enllà	,253	,507	,339	1,000	,309	,113	,278	,121
Experiència	,226	,243	,292	,309	1,000	,078	,204	,049
Inseguretat	,134	,099	,063	,113	,078	1,000	,117	,169
Influència	,178	,241	,336	,278	,204	,117	1,000	,102
Independència	,027	,103	,054	,121	,049	,169	,102	1,000

Les correlacions fluctuen en un rang similar entre les diferents preguntes. Potser les que presenten correlacions inferiors són les preguntes referides a inseguretat i independència.

En aplicar l'anàlisi corresponent en components principals, i utilitzant el criteri de valor propi superior a 1, per a l'extracció dels components, obtenim la llista següent:

Taula 8. Comunalitats

	Inicial	Extracció
Sentit de la vida	1,000	,279
Religió	1,000	,554
Obediència	1,000	,513
Més enllà	1,000	,525
Experiència	1,000	,333
Inseguretat	1,000	,562
Influència	1,000	,315
Independència	1,000	,588

Mètode d'extracció: anàlisi de components principals

Taula 9. Variància total explicada

Component	Autovalors inicials			Sumes de les saturacions al quadrat de l'extracció		
	Total	% de la variància	% acumulat	Total	% de la variància	% acumulat
1	2,554	31,926	31,926	2,554	31,926	31,926
2	1,115	13,936	45,862	1,115	13,936	45,862
3	,901	11,258	57,120			
4	,826	10,329	67,448			

Mètode d'extracció: anàlisi de components principals

Com- ponent	Autovalors inicials			Sumes de les saturacions al quadrat de l'extracció		
	Total	% de la variància	% acu- mulat	Total	% de la variància	% acumulat
5	,787	9,840	77,288			
6	,739	9,234	86,522			
7	,632	7,906	94,428			
8	,446	5,572	100,000			

Mètode d'extracció: anàlisi de components principals

Taula 10. Matriu de components^a

	Component	
	1	2
Sentit de la vida	,526	-,046
Religió	,735	-,115
Obediència	,685	-,208
Més enllà	,723	-,056
Experiència	,558	-,146
Inseguretat	,267	,701
Influència	,560	,044
Independència	,221	,734

Mètode d'extracció: anàlisi de components principals. a. 2 components extrets

Dels vuit components possibles, solament els dos primers compleixen el criteri d'autovalor superior a 1, encara que un tercer factor està a punt d'arribar a aquest límit ($\lambda_3 = 0,901$). En tot cas, s'extreuen dos components.

L'última de les taules presentades (matriu de components) ens mostra les saturacions o correlacions entre els diferents ítems i els dos components.

Recordem que la comunalitat dels ítems representa la variabilitat explicada de l'ítem pels factors extrets. Així, en l'ítem 1 ("Sentit de la vida"):

$$h_1^2 = (0,526)^2 + (-0,046)^2 = 0,279$$

Explicaria un 27,9% de la variabilitat de l'ítem. És la més baixa de tots els ítems del test. Potser aquest ítem saturaria amb un tercer component.

El valor propi (autovalor) de cada component es calcula amb la suma de quadrats de les correlacions (saturacions) entre ítems i component.

Component 1: $\lambda_1 = (0,526)^2 + (0,735)^2 \dots + (0,221)^2 = 2,554$

Component 2: $\lambda_2 = (-0,046)^2 + (-0,115)^2 + \dots + (0,734)^2 = 1,115$

El primer component ($2,554/8 = 0,319$) explicaria un 31,9% de la variabilitat presentada pels ítems, mentre que el segon ($1,115/8 = 0,139$) n'explicaria un 13,9%. Un 45,8% de la variància total és explicada per la combinació dels dos factors o components.

L'anàlisi de la matriu de saturacions ens permetrà intentar buscar una interpretació als dos components, en funció dels ítems que correlacionin de manera més alta.

En l'exemple observem que el primer component té saturacions elevades en els ítems 1, 2, 3, 4, 5, i 7; mentre que les correlacions són baixes en els ítems 6 i 8. En el segon component s'esdevé just el contrari: té correlacions altes amb els ítems 6 i 8 i, en canvi, baixes en els altres.

De vegades, a simple vista la solució final no presenta una interpretació tan fàcil. En aquests casos és possible utilitzar una rotació dels eixos per aconseguir que les correlacions siguin fortes en un eix o component i baixes en els altres. Recordem que els eixos són ortogonals i no estan relacionats entre ells. Una de les tècniques matemàtiques de rotació dels eixos és la rotació varimax, que és la més utilitzada en processos exploratoris, encara que els programes estadístics n'incorporen d'altres com les rotacions quartimax, equimax, etc.

La matriu de saturacions amb la solució de les rotacions amb el mètode varimax, amb les dades de l'exemple, quedaria com segueix:

Taula 11. Matriu de components rotats^a

	Component	
	1	2
Sentit de la vida	,522	,080
Religió	,742	,062
Obediència	,715	-,040
Més enllà	,715	,117
Experiència	,577	-,010
Inseguretat	,093	,744
Influència	,533	,175
Independència	,041	,765

Mètode d'extracció: anàlisi de components principals. Mètode de rotació: normalització varimax amb Kaiser. a. La rotació ha convergit en 3 iteracions.

Veiem com la solució de les rotacions confirma en aquest cas la conclusió prèvia, sis de les preguntes saturen en el primer component. Veient els sis ítems que saturen aquest component –“Sentit de la vida”, “Religió”, “Obediència”, “Més enllà”, “Experiència”, “Influència”–, podem interpretar aquest component com la mesura de l’actitud sobre idees religioses, que era el motiu inicial de l’elaboració del test.

El segon component, se suposa que no esperat inicialment en l’elaboració del qüestionari, satura els ítems “Inseguretat” i “Independència”. Aquest segon component es pot interpretar com una mesura de l’actitud envers la dependència personal.

Com que només es tracta de dos components, és possible representar gràficament en un espai bidimensional els dos eixos que representen els components i la situació dels ítems respecte a aquests en funció de les saturacions obtingudes.

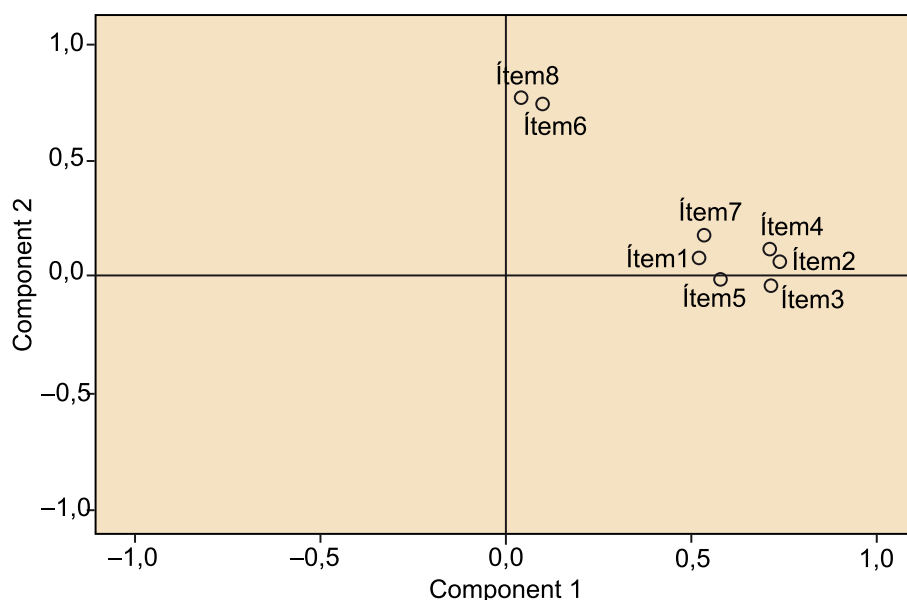


Figura 5. Gràfic de components en espai amb rotació

El gràfic dóna una informació visual ràpida de l’agrupació dels ítems en els dos components.

Una vegada obtinguts els factors terminals, podem calcular els valors o puntuacions factorials de les dimensions teòriques que hem associat als factors respectius, idees religioses (component 1) i dependència (component 2).

Taula 12. Matriu de coeficients per al càlcul de les puntuacions en els components

	Component	
	1	2
Sentit de la vida	,210	,009

Mètode d’extracció: anàlisi de components principals. Mètode de rotació: normalització varimax amb Kaiser.

	Component	
	1	2
Religió	,304	-,032
Obediència	,305	-,118
Més enllà	,287	,018
Experiència	,243	-,076
Inseguretat	-,047	,635
Influència	,204	,090
Independència	-,072	,660

Mètode d'extracció: anàlisi de components principals. Mètode de rotació: normalització varimax amb Kaiser.

Per a cada subjecte i és possible calcular les puntuacions factorials de les noves variables *Idees religioses* i *Dependència*:

$$Idees religioses_i = 0,21 \cdot Zi_{item1} + 0,304 \cdot Zi_{item2} + \dots + 0,204 \cdot Zi_{item7} - 0,072 \cdot Zi_{item8}$$

$$Dependència_i = 0,009 \cdot Zi_{item1} - 0,032 \cdot Zi_{item2} + \dots + 0,09 \cdot Zi_{item7} + 0,66 \cdot Zi_{item8}$$

Els diferents paquets estadístics permeten el càlcul automàtic i la generació d'aquestes noves variables generades i que representarien la mesura dels constructes subjacents.

5. Evidència de validesa basada en la relació amb altres variables

5.1. Concepte

En el procés de validació d'una nova prova psicològica ens podem ajudar d'altres instruments de mesura del constructe d'interès que estiguin contrastats i considerats vàlids i fiables. En aquest procés parlarem de validesa convergent, o correlació entre puntuacions del test amb altres mesures del mateix constructe fetes a partir de diferents tècniques o indicadors.

Les diferents tècniques estadístiques de relació entre variables ens serviran per a determinar el coeficient de validació entre les dues variables. Així, el més utilitzat serà el coeficient de correlació de Pearson, en cas que les dues variables siguin quantitatives, però també qualsevol de les seves variacions, com els coeficients de Spearman, biserial puntual, biserial, phi, tetracòrica, etc., en funció de com siguin les dues variables que s'han de relacionar.

Taula 13

Variable A	Variable B	Coefficient correlació
Numèrica (interval o raó)	Numèrica (interval o raó)	r de Pearson
Numèrica (interval o raó)	Numèrica (ordinal)	r_s Spearman o τ de Kendall
Numèrica (ordinal)	Numèrica (ordinal)	r_s Spearman o τ de Kendall
Qualitativa	Qualitativa	V de Cramer
Qualitativa (dicotòmica)	Numèrica (interval o raó)	r_b biserial o r_{bp} biserial puntual
Qualitativa (dicotòmica)	Qualitativa (dicotòmica)	Φ phi o r_t tetracòrica

Per poder conèixer el nivell de benestar físic i psicològic en gent gran, ens interessa validar un nou qüestionari que hem elaborat, per determinar el grau d'independència en les activitats bàsiques de la vida diària (ABVD). Amb aquest objecte podem utilitzar algunes de les proves que ja hi ha en el mercat i que es troben prou contrastades. En una mostra de tres-cents subjectes més grans de setanta anys, i ingressats en centres geriàtrics, administrem la nova prova elaborada conjuntament amb l'escala de mesura d'independència funcional (MIF o FIM) Keith, Granger, Hamilton i Sherwin, 1987) i l'escala de grau d'autonomia de Barthel (Mahoney i Barthel, 1965).

En la taula següent es mostra la matriu de correlacions de Pearson (dades simulades) entre les tres proves administrades:

Taula 14

	ABVD	FIM	Barthel
ABVD	1		
MIF	0,69	1	
Barthel	0,67	0,77	1

Els valors de la correlació entre la nova prova (ABVD) i les escales MIF i Barthel presenten valors prou alts (0,69 i 0,67, respectivament), la qual cosa indica una validesa concurrent alta de la nova prova elaborada amb les tècniques prèvies per a mesurar el grau d'autonomia de les persones grans analitzades.

5.2. Evidència de decisió (sensibilitat i especificitat)

En situacions en què la prova feta tingui com a objectiu el diagnòstic o classificació dels subjectes en dos grups (diagnòstic negatiu - diagnòstic positiu), parlarem de la validesa de decisió, quan comparem aquesta nova prova amb un altre mètode de diagnòstic anterior prou contrastat. En la validesa de decisió podem distingir dos processos: d'una banda, la sensibilitat de la prova, definida com la capacitat d'aquesta per a detectar veritables positius; i, d'altra banda, l'especificitat, com la capacitat per a determinar diagnòstics negatius veritables.

Taula 15

		Diagnòstic prova inicial		
		Positiu	Negatiu	Total
Diagnòstic nova prova	Positiu	Decisió correcta (f_{11})	Fals positiu (f_{12})	$f_{1.}$
	Negatiu	Fals negatiu (f_{21})	Decisió correcta (f_{22})	$f_{2.}$
	Total	$f_{.1}$	$f_{.2}$	n

Una mesura de l'acord absolut per mitjà de les dues proves diagnòstiques consistirà a calcular la proporció d'acord (P_c) entre totes dues tècniques, a partir de la raó entre la suma de decisions correctes i el total de decisions.

$$P_c = \frac{f_{11} + f_{22}}{n} \cdot 100$$

La sensibilitat de la nova prova l'obtindrem a partir de la proporció de subjectes classificats correctament com a veritables positius.

$$\text{Sensibilitat} = \frac{\text{Diagnòstics positius de la prova}}{\text{Total diagnòstics positius}} = \frac{f_{11}}{f_{.1}}$$

Mentre que l'especificitat s'obté mitjançant el quocient dels diagnosticats sense trastorn per la prova entre el total de diagnòstics negatius.

$$\text{Especificitat} = \frac{\text{Diagnòstics negatius de la prova}}{\text{Total diagnòstics negatius}} = \frac{f_{22}}{f_{.2}}$$

"Tot i que es fa difícil establir uns valors que es puguin considerar acceptables per a aquests dos indicadors, en general, es considera que una prova té una exactitud acceptable quan la seva sensibilitat i especificitat superen el llindar de 0,8 (Domènech i Granero, 2003)."

Un índex global per a valorar la validesa, el proporciona el càlcul del coeficient kappa, establert inicialment com a indicador de l'acord entre dos observadors. L'avantatge que presenta és que és fàcil d'interpretar, similar a la d'altres indicadors de relació entre variables. En efecte, el seu valor fluctua entre 0 (cap acord) i 1 (màxim acord). La fórmula de càlcul és senzilla:

$$K = \frac{F_c - F_a}{n - F_a}$$

en què

$$F_c = f_{11} + f_{22} \quad y \quad F_a = \frac{f_{1.} \cdot f_{.1} + f_{2.} \cdot f_{.2}}{n}$$

Taula 16. Criteris Alman per a interpretar kappa

Valor	Relació
0-0,20	Inexistent
0,21-0,40	Molt baixa
0,41-0,60	Moderada
0,61-0,80	Bona
0,81-1	Intensa

En una consulta psicològica es vol validar una nova prova, més simple que les tradicionals, per al diagnòstic del trastorn de depressió dels pacients atesos. En una mostra de cinc-cents pacients atesos al centre s'administren dues proves (tradicional i versió breu) per al diagnòstic del trastorn de depressió.

Taula 17

		Diagnòstic tradicional depressió		
		Positiu	Negatiu	Total
Versió breu escala Hamilton	Positiu	125	50	175
	Negatiu	25	300	325
	Total	150	350	500

$$P_c = \frac{125+300}{500} = 0,85$$

Les dues proves presenten un percentatge d'acord (85%) elevat.

Així mateix, els valors de sensibilitat i especificitat indiquen que la nova prova té una bona capacitat per a detectar subjectes amb trastorn depressiu (sensibilitat = 0,83), i per a detectar subjectes sense depressió (especificitat = 0,86).

$$\text{Sensibilitat} = \frac{125}{150} = 0,83$$

$$\text{Especificitat} = \frac{300}{350} = 0,86$$

$$F_c = 125 + 300 = 425 \quad \text{i} \quad F_a = \frac{175 \cdot 150 + 325 \cdot 350}{500} = 280$$

$$K = \frac{425 - 280}{500 - 280} = 0,66$$

El càlcul de l'índex kappa d'acord entre les dues proves ($K = 0,66$) indica una bona relació entre aquestes. Per tant, sembla adequat pensar que l'economia de temps i esforç (tant per al pacient com per al terapeuta) justificaria la nova escala de diagnòstic de depressió, en funció dels resultats obtinguts en la validesa de decisió.

5.3. Evidències convergents i/o discriminants

Fins ara, en aquest apartat, ens hem referit a escales que volen mesurar un sol constructe psicològic. Si ens referim a proves formades per ítems que mesuren diferents constructes (trets múltiples), podrem diferenciar dos tipus de validesa. D'una banda, la validesa convergent (enunciada anteriorment), o sigui, la validesa que determinen diferents proves que mesuren el mateix constructe, i, d'altra banda, la validesa discriminant, que està determinada per la mesura de diferents constructes dins de la mateixa prova.

La matriu de correlacions entre les puntuacions dels diferents trets obtinguts a partir de les diferents escales (matriu multitret-multimètode; Campbell i Fiske, 1959) ens servirà per a determinar els diferents valors de validesa convergent i discriminant.

Imaginem la situació en què disposem de tres escales diferents, formades per ítems que mesuren els mateixos dos constructes subjacents: escales *A*, *B* i *C*, que mesuren els constructes 1 i 2. Per a cada subjecte analitzat tindrem, per tant, sis puntuacions diferents obtingudes per la combinació de cada prova i cada tret analitzat.

Taula 18

	A1	A2	B1	B2	C1	C2
A1	Fi					
A2	Vd	Fi				
B1	Vc		Fi			
B2		Vc	Vd	Fi		
C1	Vc		Vc		Fi	
C2		Vc		Vc	Vd	Fi

En la taula anterior, les diferents escales es representen per les lletres *A*, *B* i *C*, i dins de cada escala 1 i 2 representen els dos constructes que es volen analitzar.

Si observem la matriu multitret-multimètode (MTMM), trobem en la diagonal principal valors de fiabilitat de les escales (valors 1 si s'obtenen en una única administració).

Els valors de validesa convergent es troben en les combinacions dels mateixos trets i diferents escales (per exemple, casella *A1* i *B1*), esperant que aquests valors de validesa siguin prou alts, la qual cosa indica convergència de les diferents maneres de mesurar el constructe i aporta evidència real de l'existència del constructe.

Els valors de validesa discriminant seran els coeficients de correlació obtinguts dins de la mateixa escala per les puntuacions dels diferents trets. En aquest cas, esperem que els diferents constructes siguin prou independents entre ells perquè les correlacions siguin properes a 0.

En una mostra de sis-cents subjectes s'han utilitzat tres proves diferents de personalitat (tests 1, 2 i 3), formades cada una per ítems referits als tres mateixos constructes de la personalitat (trets *A*, *B* i *C*).

La taula següent mostra els valors de la matriu de correlacions entre les nou variables obtingudes de la combinació de les tres proves i els tres trets (3×3).

Taula 19

	A1	B1	C1	A2	B2	C2	A3	B3	C3
A1	1								
B1	0,03	1							
C1	0,28	0,17	1						
A2	0,73	0,14	0,22	1					
B2	0,15	0,69	0,03	0,18	1				
C2	0,12	0,04	0,81	0,03	0,15	1			

	A1	B1	C1	A2	B2	C2	A3	B3	C3
A3	<u>0,77</u>	0,11	0,09	<u>0,77</u>	0,21	0,16	1		
B3	0,21	<u>0,75</u>	0,18	0,21	<u>0,68</u>	0,03	<i>0,15</i>	1	
C3	0,19	0,05	<u>0,78</u>	0,22	0,09	<u>0,72</u>	<i>0,14</i>	<i>0,09</i>	1

Els valors de validesa convergent són els valors destacats amb el subratllat. Els valors de validesa discriminant es destaquen amb la cursiva.

Si ens fixem en el tret B, es detecta convergència a partir de la mesura per mitjà de diferents proves:

$$r(B1 - B2) = 0,69$$

$$r(B1 - B3) = 0,75$$

$$r(B2 - B3) = 0,68$$

Si ens fixem en l'escala 1, observem que hi ha prou independència entre les diferents mesures dels tres trets mesurats:

$$r(A1 - B1) = 0,03$$

$$r(A1 - C1) = 0,28$$

$$r(B1 - C1) = 0,17$$

5.4. Evidències basades en les relacions test-criteri

De vegades, un test o prova psicològica construït per a mesurar un determinat constructe psicològic pot estar relacionat amb una altra variable d'interès que es denomina *criteri*.

Per exemple, imaginem que hem elaborat una prova vàlida que ens permet mesurar la capacitat de raonament numèric de les persones (test), i observem que presenta una correlació molt alta amb els resultats que obtenen els subjectes en una determinada prova de matemàtiques (criteri).

Podem distingir tres tipus de situacions:

- Validesa concurrent o simultània
- Validesa predictiva
- Validesa retrospectiva

5.4.1. Validesa concurrent o simultània

En aquest cas el test i el criteri es mesuren de manera simultània. Obtindrem validesa concurrent en obtenir valors alts de coeficients de correlació entre les puntuacions del test i del criteri. Per tant, ens permet validar el test, inicialment elaborat per a mesurar una altra variable, per a mesurar el criteri.

En funció del tipus d'escala de mesura utilitzat tant per a les puntuacions del test com del criteri, utilitzarem un tipus o un altre de coeficient per a mesurar la correlació.

Com a exemple, vegem la situació següent, en la qual s'ha administrat un test de raonament numèric a un grup de vint subjectes just abans de fer una determinada prova de matemàtiques:

Taula 20

Test raonament	Nota matemàtiques
100	10
100	9
100	9
90	8
90	8
80	9
70	7
70	7
70	7
70	5
70	4
70	4
50	3
40	3
40	2
40	2
30	1
20	1
20	3
20	2

Una visió del gràfic de dispersió ens donarà una idea de si hi ha relació lineal entre totes dues proves, o no:

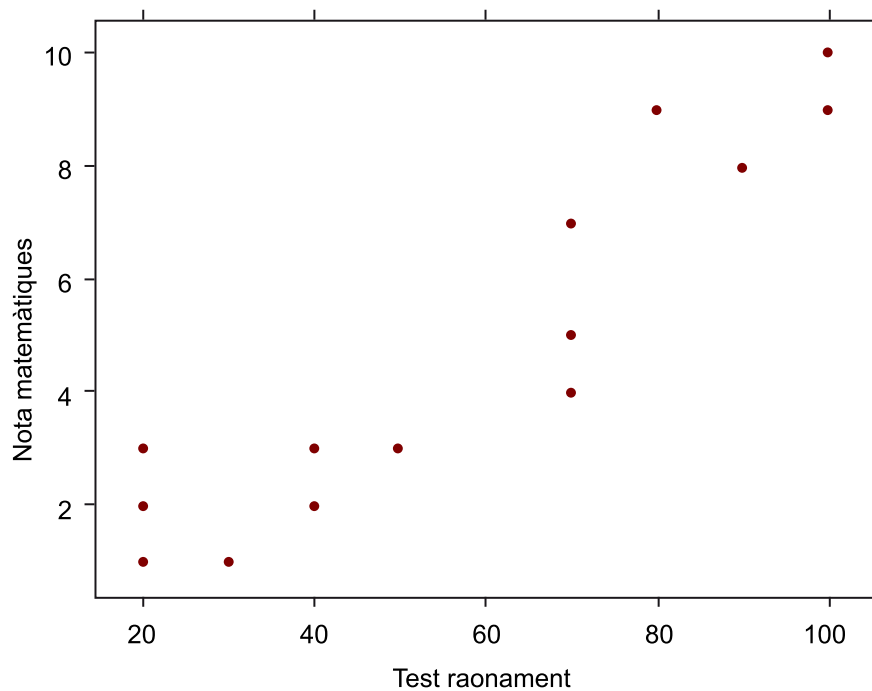


Figura 6. Gràfic de dispersió (núvol de punts)

En calcular el coeficient de correlació de Pearson:

```
> rcorr.adjust(Dades[,c("Nota.Matemàtiques", "Test.Raonament")], type="pearson")
```

	Nota. Matemàtiques	Test.Raonament
Nota. Matemàtiques	1.00	0.92291
Test.Raonament	0.92291	1.00

n= 20

S'obté un valor de validesa concurrent igual a 0,92291, que indica una forta relació directa i propera a 1.

Una mesura de la bondat de l'ajust lineal entre les dues variables es defineix per r^2 , o sigui el valor del quadrat de la correlació, en el nostre exemple $r^2 = (0,92291)^2 = 0,8518$. Aquest valor, que es denomina *coeficient de determinació*, multiplicat per 100, indica el percentatge de variabilitat en la variable criteri, que és explicat per la relació amb la variable independent. Per tant, el 85,18% de la variabilitat que presenten les puntuacions obtingudes en la nota de matemàtiques estaria explicada per la relació que presenta amb els valors que s'han obtingut en el test de raonament numèric.

5.4.2. Validesa predictiva

Si sabem que un determinat test i una variable criteri es troben altament relacionats, serà possible utilitzar els valors obtinguts en el test per a la predicció o pronòstic dels valors que s'obtingran en el criteri. Parlarem en aquest cas de la validesa predictiva que té el test respecte a la variable criteri.

Per exemple, si volem seleccionar un candidat per a un determinat lloc de treball, podem utilitzar determinades proves que tinguin una alta validesa predictiva respecte al rendiment futur dels candidats en el lloc de treball. O, utilitzant l'exemple esmentat més amunt, podem fer un pronòstic de la nota que obtindran els subjectes en una determinada prova de matemàtiques a partir de les puntuacions que van obtenir al seu moment en el test de raonament numèric.

Quan tant les puntuacions del test com el criteri són puntuacions numèriques, i hem calculat el coeficient de correlació de Pearson corresponent, que és significatiu estadísticament, és possible establir un model de regressió lineal per a poder fer el pronòstic dels valors del criteri.

Les puntuacions en el test constitueixen la variable independent del model (variable predictora), mentre que el criteri representa la variable dependent.

Regressió lineal simple

És el cas més senzill i solament disposem d'una variable independent (X) i una variable dependent (Y).

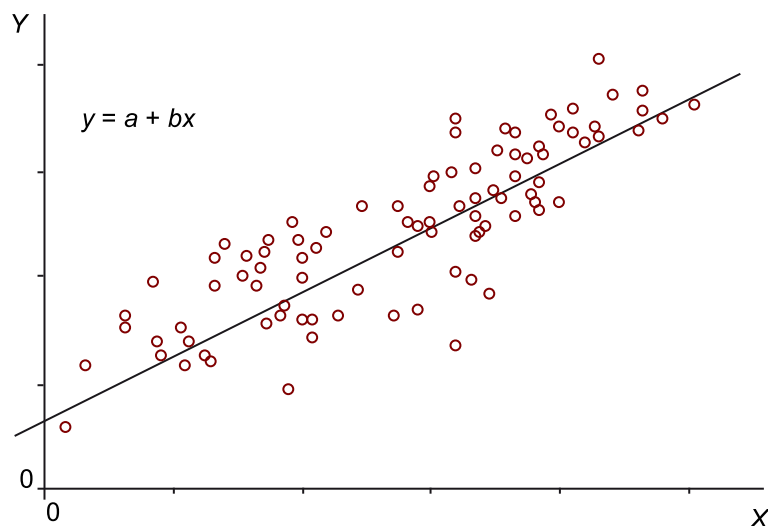


Figura 7. Gràfic de dispersió amb recta de regressió

La regressió lineal descriu una relació lineal entre Y i X , això és, en el gràfic de dispersió representa la recta que millor s'ajusta al núvol de punts.

L'expressió d'una línia recta és $y = a + bx$, en què b representa el pendent de la recta, o sigui, el canvi que es produeix en y a partir del canvi que es produeixi en x , i a es denomina *intersecció* o *intercepta*, i y és el valor que pren y quan x és igual a 0.

Per a trobar els coeficients de la regressió, a i b , usem un mètode d'estimació molt conegut en estadística, el mètode de mínims quadrats, el qual minimitza la suma dels quadrats de les diferències (o residus) entre els valors y_i i els valors estimats segons la recta de regressió $y'_i = a + bx_i$.

A partir de les dades (x_i, y_i) , $i = 1, \dots, n$, estimem els coeficients a i b de la recta de regressió. Així doncs, tenim:

- Pendent:

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2}$$

Lectura de la fórmula

S_{xy} : covariància entre x i y .

S_x^2 : variància de x .

- Intersecció:

$$a = \bar{y} - b \cdot \bar{x}$$

Comparant les fórmules del pendent b i del coeficient de correlació r , tenim la relació següent:

$$b = r_{xy} \cdot \frac{S_y}{S_x}$$

És possible verificar o validar el model de regressió a partir del coeficient de determinació r^2 . Recordem que més amunt l'hem definit com una mesura de bondat d'ajust, o mesura de la proximitat dels punts a la recta estimada. Representa la proporció de variància de la variable dependent explicada per la recta de regressió.

El valor $1 - r^2$ quantifica la proporció de variància que no és explicada per la regressió. A partir d'aquests dos valors podem calcular un estadístic de contrast:

$$F_{EC} = \frac{r^2}{(1-r^2)/n-2}$$

Aquest estadístic de contrast F es distribueix seguint una distribució F de Snedecor, amb 1 grau de llibertat en el numerador i $n - 2$ graus de llibertat en el denominador.

Les hipòtesis que s'han de contrastar seran:

- H_0 : el model no és vàlid, no hi ha relació.
- H_1 : sí que hi ha relació, per tant el model sí que és vàlid.

Seguint amb l'exemple del test de raonament numèric i el criteri de la nota de matemàtiques, el resultat amb el programa *R* és el següent:

```
> summary(RegModel.1)
```

```
Call:
```

```
lm(formula = Nota.Matemàtiques ~ Test.Raonament, data = Dades)
```

```

Residuals:
    Min       1Q   Median       3Q      Max
-2.01102 -0.96970 -0.03857  0.98898  2.05785

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.085399   0.673899  -1.611   0.125
Test.Raonament  0.101377   0.009969  10.170 6.89e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.201 on 18 degrees of freedom

Multiple R-squared:  0.8518, Adjusted R-squared:  0.8435
F-statistic: 103.4 on 1 and 18 DF, p-value: 6.89e-09

```

Un cop consultada la sortida del programa, podem especificar el model de regressió:

$$\text{Nota matemàtiques} = -1,085 + 0,1014 \times \text{Test} + \text{Residual}$$

També observem que el model està verificat, ja que el valor de p (grau de significació) que acompanya el valor de l'estadístic de contrast ($F = 103,4$) és tendent a 0. Per tant, res no s'oposa a rebutjar la hipòtesi nul·la, i el model està validat.

Si representem la recta de regressió en el gràfic de dispersió:

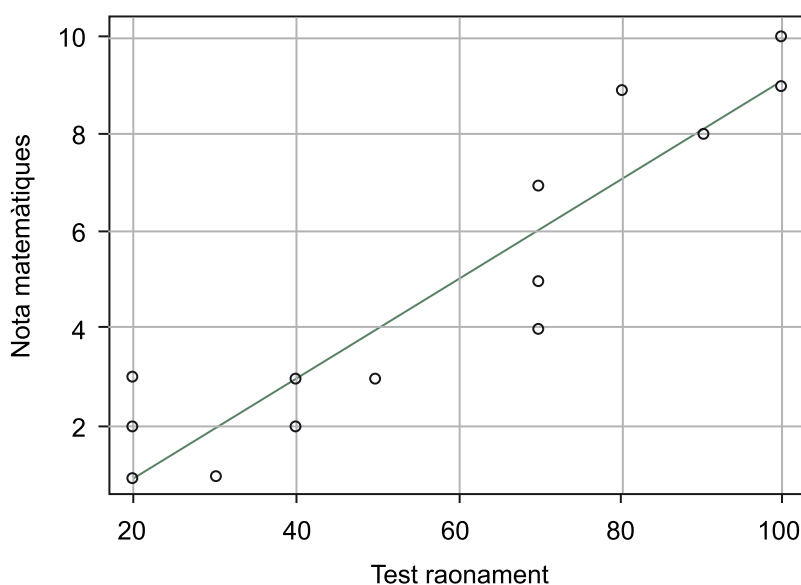


Figura 8. Gràfic de dispersió amb ajust de la recta

A partir de l'expressió de la recta de regressió, podem fer el pronòstic, per a cada subjecte, del valor de la nota de matemàtiques en funció del test, com també el càlcul del residual mitjançant la diferència entre la puntuació real obtinguda i la puntuació pronosticada.

Taula 21

Test raonament	Nota matemàtiques	Pronòstic	Residual
100	10	9,0546	0,9454
100	9	9,0546	-0,0546
100	9	9,0546	-0,0546
90	8	8,0406	-0,0406
90	8	8,0406	-0,0406
80	9	7,0266	1,9734
70	7	6,0126	0,9874
70	7	6,0126	0,9874
70	7	6,0126	0,9874
70	5	6,0126	-1,0126
70	4	6,0126	-2,0126
70	4	6,0126	-2,0126
50	3	3,9846	-0,9846
40	3	2,9706	0,0294
40	2	2,9706	-0,9706
40	2	2,9706	-0,9706
30	1	1,9566	-0,9566
20	1	0,9426	0,0574
20	3	0,9426	2,0574
20	2	0,9426	1,0574

```
> numSummary(Dades[,c("Nota.Matemàtiques", "Pronòstic", "Residual")],
+ statistics=c("mean", "sd", "var"))
```

	mean	sd	var	n
Nota.Matemàtiques	5.2000	3.036619	9,221	20
Pronòstic	5.2014	2.803138	7,855	20
Residual	-0.0014	1.169176	1,367	20

Tal com hem indicat més amunt, el coeficient de determinació es defineix com el quocient entre la variància explicada per la regressió i la variància total de la variable criteri:

$$r^2 = \frac{s_{y'}^2}{s_y^2} = 1 - \frac{s_{y-y'}^2}{s_y^2}$$

Amb les dades del nostre exemple:

$$r^2 = \frac{7,855}{9,221} = 1 - \frac{1,367}{9,221} = 0,8518$$

Per tant, a partir dels valors del coeficient de determinació i de la variància de la variable criteri, és possible, aïllant l'expressió, obtenir el valor de la variància dels errors:

$$s_{y-y'}^2 = s_y^2(1 - r^2)$$

La desviació típica dels errors o error típic o estàndard de l'error ens ajudarà en l'estimació per interval de nous valors desconeguts.

$$s_{y-y'} = s_y \sqrt{1 - r^2}$$

En efecte, si necessitem fer un pronòstic en el criteri a partir d'un nou valor en el test, o variable predictora, és possible fer-lo de manera puntual, però, donada una probabilitat, aconseguirem estimacions millors si es fa per interval.

$$IC^{1-\alpha} \rightarrow y' \pm t_{n-1;\alpha/2} \cdot s_{y-y'}$$

Imaginem que un nou subjecte obté una puntuació igual a 60 en el test de raonament numèric, l'estimació puntual del valor en la nota de matemàtiques serà:

$$Nota matemàtiques = -1,085 + 0,1014 \cdot 60 = 4,999$$

Si fem una estimació per interval, amb un nivell de confiança del 95%:

$$IC^{0,95} \rightarrow 4,999 \pm 2,093 \cdot 1,169 = [2,552 \quad 7,446]$$

Amb una probabilitat de 0,95, el valor en el criteri d'un subjecte que obtingui una puntuació de 60 en el test estarà entre 2,552 i 7,446 punts.

Regressió lineal múltiple

El model lineal general planteja que una variable dependent (criteri) sigui funció de diverses variables independents, situació, d'altra banda, força més habitual. En el cas de dues variables independents, l'expressió que relaciona les

Lectura de la fórmula

s_y^2 : variància de la variable dependent Y .

$s_{y'}^2$: variància dels pronòstics obtinguts mitjançant l'equació de regressió.

$s_{y-y'}^2$: variància dels errors produïts.

Lectura de la fórmula

$1 - \alpha$: nivell de confiança de l'interval construït.

t : valor de la distribució t de Student-Fisher tabulat en funció de α i dels graus de llibertat ($n - 1$).

tres variables serà la fórmula d'un pla, en les situacions en què hi hagi més de dues variables independents, situacions multivariants, parlarem de l'hiperplà de regressió.

$$Criteri = a + b_1 \cdot Test_1 + b_2 \cdot Test_2 + b_3 \cdot Test_3 + \dots + b_p \cdot Test_p + Residual$$

Recuperant l'exemple anterior, imaginem que als vint subjectes als quals s'ha administrat un test de raonament juntament amb un test de càlcul mental, abans de fer una prova de matemàtiques, que representa el criteri que més endavant volem pronosticar.

Taula 22

Test raonament	Test càlcul	Nota matemàtiques
100	9	10
100	8	9
100	8	9
90	8	8
90	7	8
80	9	9
70	6	7
70	5	7
70	6	7
70	6	5
70	4	4
70	4	4
50	5	3
40	4	3
40	4	2
40	5	2
30	3	1
20	2	1
20	3	3
20	3	2

En calcular la matriu de correlacions de Pearson:

Taula 23

	Test raonament	Test càlcul	Nota matemàtiques
Test raonament	1		
Test càlcul	0,891763351	1	
Nota matemàtiques	0,922905961	0,925261006	1

Observem que el test de càlcul mental també està altament correlacionat amb la variable criteri (nota de matemàtiques). Així mateix, veiem una alta correlació entre les dues variables independents, test de raonament numèric i test de càlcul mental ($r = 0,89176$).

L'anàlisi de regressió múltiple ens ajudarà a determinar si la incorporació d'aquesta nova variable augmenta, significativament, la variabilitat explicada per la regressió en la variable criteri.

L'anàlisi de regressió es basarà en l'anàlisi de la relació conjunta entre la variable criteri i el conjunt de les dues variables independents. El quadrat d'aquesta correlació múltiple serà el nou coeficient de determinació.

La verificació del model es farà a partir de l'expressió:

$$F_{EC} = \frac{r^2 / p}{(1-r^2) / (n-p-1)}$$

Lectura de la fórmula

p és igual al nombre de variables independents en el model.

L'estadístic de contrast F es distribueix seguint una distribució F de Snedecor, amb p graus de llibertat en el numerador i $(n - p - 1)$ graus de llibertat en el denominador.

A continuació es presenta la sortida del programa obtinguda en aquest exemple mitjançant l'ús del programa R:

```
> RegModel.2 <- lm(Nota.Matemàtiques~Test.Numèric+Test.Raonament, data=Dades)

> summary(RegModel.2)

Call:
lm(formula = Nota.Matemàtiques ~ Test.Numèric + Test.Raonament, data = Dades)

Residuals:
    Min       1Q   Median       3Q      Max
-1.72686 -0.64006 -0.00108  0.52167  1.74005

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
                
```

```

(Intercept)      -1.91579  0.62646   -3.058    0.00711 **
Test.Càlcul       0.70883  0.23719    2.988    0.00825 **
Test.Raonament    0.05246  0.01835    2.858    0.01088 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.001 on 17 degrees of freedom
Multiple R-squared:  0.9028, Adjusted R-squared:  0.8914
F-statistic: 78.96 on 2 and 17 DF, p-value: 2.481e-09

```

El valor de coeficient de determinació és 0,9028, per tant un 90,28% de la variància de la variable criteri s'explica per la regressió entre aquesta variable i la combinació de les dues variables independents.

L'equació del pla de regressió es troba verificada globalment, ja que el valor de l'estadístic de contrast ($F = 78,96$) indicia una probabilitat tendent a 0 (p value) que no hi hagi relació entre les variables.

També és important observar si es troben significats en els diferents coeficients de la regressió (b). En aquest cas, tant el coeficient que afecta el test de raonament (p value = 0,01) com el que afecta el test de càlcul mental ($p = 0,008$) estan verificats, ja que els graus respectius de significació associats són propers a 0.

Per tant, l'especificació de l'equació resultant del model de regressió quedarà de la manera següent:

Nota matemàtiques = $-1,916 + 0,709 \cdot \text{Test_càlcul} + 0,052 \cdot \text{Test_raonament} + \text{Residual}$

Així mateix, la sortida del programa informa del valor de l'error típic o estàndard de l'error ($s_{y-y'} = 1,001$), valor necessari per a l'estimació per interval dels valors del criteri. En efecte, per a un nou subjecte que obtingués una puntuació igual a 6 en el test de càlcul mental i 60 en el test de raonament numèric, l'estimació puntual del valor en la nota de matemàtiques serà:

$$\text{Nota matemàtiques} = -1,916 + 0,709 \cdot 6 + 0,052 \cdot 60 = 5,458$$

Si fem una estimació per interval, amb un nivell de confiança del 95%:

$$IC^{0,95} \rightarrow 5,458 \pm 2,093 \cdot 1,001 = [3,363 \quad 7,553]$$

Amb una probabilitat de 0,95, el valor en el criteri d'un subjecte que obtingui una puntuació de 6 en la prova de càlcul i de 60 en el test de raonament estarà entre 3,363 i 7,553 punts.

Altres tècniques estadístiques d'anàlisi multivariable

En funció del tipus d'escala de mesura utilitzada per a les variables criteri i les variables predictores, serà necessari aplicar alguna de les diferents tècniques d'anàlisi de dades multivariables.

Per exemple, si disposem d'una variable criteri mesurada en escala nominal, i per tant de tipus qualitatiu o categòric, mentre que les variables independents són de tipus quantitatiu o numèric, podem utilitzar una tècnica de classificació, com l'anàlisi discriminant. En efecte, suposem que la variable criteri nota de matemàtiques, de l'exemple que hem utilitzat, estigués codificada en suspens, aprovat, notable, excel·lent, o que els subjectes estiguessin directament simplement dividits entre aprovats i suspesos.

Aplicar una anàlisi discriminant ens permetria determinar la millor funció discriminant que aconseguixi la classificació dels subjectes, en funció de les puntuacions obtingudes en les proves predictores del càlcul mental i el raonament numèric. La funció discriminant establirà l'estimació dels pesos (coeficients) i la combinació lineal de les variables independents (discriminants), de manera que els grups siguin, des del punt de vista estadístic, tan diferents com sigui possible.

Una altra opció pot ser que tant la variable criteri com les variables independents estiguin codificades en categories. En aquest cas seria aplicable la tècnica del model lògit. Aquest model, basat en els models lineals logarítmics, pretén, seguint l'enfocament de la regressió múltiple, trobar l'expressió d'associació entre la variable criteri i les variables independents, tenint en compte també la interacció entre les variables independents.

5.4.3. Validesa retrospectiva

La validesa concurrent entre un o diversos tests i el criteri, que pot ser útil per a la predicció futura de la variable criteri, en certes situacions i donades certes conseqüències mesurades per mitjà del criteri, també pot servir per a trobar les causes als valors obtinguts.

En aquest cas, la variable criteri s'ha registrat abans de les variables predictores. Per exemple, en psicologia és habitual aplicar diferents proves psicològiques que permetin donar una explicació a una determinada conducta d'un subjecte.

5.5. Generalització de la validesa

El concepte de *generalització de la validesa* es refereix al fet d'estendre la validesa establerta entre test i criteri a altres situacions o a grups de subjectes diferents dels utilitzats inicialment en el càlcul.

Per exemple, imaginem que es troben validades determinades proves per a la selecció correcta de personal per a determinats llocs d'administratiu en un banc *A*. Si és possible utilitzar les mateixes proves per a la selecció d'administratius en un altre banc *B*, podrem considerar que la validesa s'ha generalitzat. En aquest cas es tractaria de subjectes diferents dels utilitzats inicialment.

Un altre cas podria ser que les proves de selecció per a administratius de banca s'utilitzessin per a seleccionar personal administratiu en companyies d'assegurances. En aquest cas parlariem de generalització també a situacions diferents.

6. Evidència de validesa basada en les conseqüències de l'aplicació

Quan es prenen decisions a partir de l'aplicació d'un test, i no es tracta només de descriure o interpretar sense que se'n derivin accions, s'ha de pensar en les conseqüències que té aplicar aquest qüestionari (Shepard, 1997). Els tests s'han d'usar quan es maximitzin les conseqüències positives (beneficis) i es minimitzin les negatives (costos) derivades de la seva aplicació.

Els tests s'apliquen esperant aconseguir algun tipus de benefici (poder seleccionar el millor tractament terapèutic, situar els treballadors d'una empresa en el lloc més adequat, millorar les tècniques didàctiques emprades, etc.) de la informació obtinguda. Un dels propòsits fonamentals de la validació és indicar en quins casos es poden obtenir aquests beneficis.

Dins d'aquest concepte cal diferenciar entre evidències que són rellevants per a la validesa i evidències que són importants per a les polítiques socials però se situen fora del concepte de *validesa*. Aquesta diferència es fa més important quan les conseqüències que es deriven del test són diferents per a diferents grups. Per exemple, si se sap que hi ha diferències entre homes i dones en les puntuacions d'un test emprat per a la selecció de personal, això afectarà l'ús del test, però no es podria afirmar res sobre les evidències de validesa basades en les conseqüències de l'aplicació. Per a això s'ha de fer un estudi més detallat de les conseqüències. Si les diferències es deuen al fet que l'aspecte avaluat es distribueix de manera diferent entre els grups en la població, les diferències obtingudes no impliquen que les decisions que es deriven de l'aplicació del test manquin de validesa. El problema sorgeix quan aquestes diferències es deuen al fet que s'estan valorant habilitats que no estan relacionades amb la tasca que duran a terme els seleccionats o quan el test és sensible a algunes característiques dels candidats que no es vol que estiguin relacionades amb el constructe a mesurar. En el primer cas, no es pot concloure la manca d'indicis de validesa respecte a les conseqüències si bé en les dues últimes situacions sí. Però, evidentment, les tres situacions serien inadequades dins de les polítiques socials d'igualtat de gènere (APA, 1999).

7. Factors que afecten la validesa

Tal com s'ha comentat més amunt, un dels indicis de validesa que es poden calcular (o s'han de calcular) és la correlació entre el test i un criteri aliè a aquest. Aquesta correlació es pot veure afectada per múltiples factors com són la fiabilitat de totes dues mesures, la longitud del test i la variabilitat (dispersió) de la mostra emprada per a obtenir les puntuacions. A continuació es veuran aquests aspectes.

7.1. Fórmules d'atenuació

Quan es tracta de calcular la correlació entre un test i un criteri es parteix de les puntuacions empíriques que s'han obtingut en tots dos qüestionaris. Aquestes puntuacions estan compostes, segons el model lineal de Spearman, per la puntuació veritable i l'error de mesura que és aleatori. Per tant:

$$\begin{aligned}X &= V_x + e_x \\ Y &= V_y + e_y\end{aligned}$$

Com es pot comprovar, en correlacionar les puntuacions X i Y també es correlacionen els dos errors de mesura entre ells. Aquests errors són aleatoris, la qual cosa significa que la correlació entre tots dos ha de ser igual a 0. Per això, com més importància tinguin els errors en la puntuació obtinguda (o, el que és el mateix, com més baixa sigui la fiabilitat del test i el criteri emprats) més baixa serà la correlació entre X i Y . En definitiva, es pot trobar una regla segons la qual com més fiables siguin el test i el criteri, més gran serà la correlació entre tots dos. Per saber en quin grau ho fa s'empren les fórmules d'atenuació (Spearman, 1907).

Lectura de la fórmula

X : puntuació obtinguda en el test.

V_x : puntuació veritable en el test.

e_x : error de mesura (aleatori) en el test.

Y : puntuació obtinguda en el criteri.

V_y : puntuació veritable en el criteri.

e_y : error de mesura (aleatori) en el criteri.

7.1.1. Estimació del coeficient de validesa en el supòsit en què el test i el criteri tinguin una fiabilitat perfecta

En el cas en què se suposa que el test i el criteri tenen una fiabilitat perfecta s'assumeix que els errors de mesura són iguals a 0. En aquesta situació (en què el coeficient de fiabilitat és igual a 1), la puntuació empírica (X o Y segons es tracti del test o del criteri) és igual a la veritable.

En aquest cas, la nova correlació es pot calcular mitjançant:

$$\rho_{v_x v_y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}} \sqrt{\rho_{yy'}}$$

Exemple

La correlació entre un test d'ansietat i un criteri (depressió) és de 0,63. La fiabilitat del test és de 0,79 i la del criteri de 0,90. Quina és l'estimació d'aquesta correlació si se suposa que tots dos tenen una fiabilitat perfecta?

$$\rho_{v_x v_y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}} \sqrt{\rho_{yy'}}} = \frac{0,63}{\sqrt{0,79} \sqrt{0,90}} = 0,75$$

La correlació que s'estima entre el test i el criteri si tots dos tinguessin una fiabilitat perfecta passa de 0,63 a 0,75.

Lectura de la fórmula

ρ_{xy} : coeficient de validesa obtingut en correlacionar les puntuacions del test i el criteri.

$\rho_{xx'}$: coeficient de fiabilitat del test.

$\rho_{yy'}$: coeficient de fiabilitat del criteri.

7.1.2. Estimació del coeficient de validesa en el supòsit en què el test tingui una fiabilitat perfecta

En el cas en què només el test tingui una fiabilitat perfecta (igual a 1), l'estimació de la nova correlació queda determinada per:

$$\rho_{v_x y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}$$

En l'exemple anterior, si només el test tingués la fiabilitat perfecta el resultat seria:

$$\rho_{v_x y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}} = \frac{0,63}{\sqrt{0,79}} = 0,71$$

La correlació que s'estima entre el test i el criteri en suposar que el test té una fiabilitat perfecta canvia de 0,63 a 0,71.

7.1.3. Estimació del coeficient de validesa en el supòsit en què el criteri tingui una fiabilitat perfecta

Si és el criteri el que se suposa que té una fiabilitat perfecta, l'estimació de la correlació queda determinada per:

$$\rho_{x v_y} = \frac{\rho_{xy}}{\sqrt{\rho_{yy'}}$$

En l'exemple anterior:

$$\rho_{x v_y} = \frac{\rho_{xy}}{\sqrt{\rho_{yy'}}} = \frac{0,63}{\sqrt{0,90}} = 0,66$$

La correlació que s'estima entre el test i el criteri en suposar que el test té una fiabilitat perfecta canvia de 0,63 a 0,66.

7.1.4. Estimació del coeficient de validesa en el supòsit en què s'hagi millorat tant la fiabilitat del test com la del criteri

La situació més freqüent és la situació en què es millora la fiabilitat del test, la del criteri o la de tots dos però sense arribar a 1 (aquest esdeveniment és més teòric que pràctic). A continuació es veurà com s'estima la correlació entre test i criteri quan es milloren les fiabilitats de tots dos. Posteriorment es tractaran les altres opcions.

$$\rho_{XY} = \frac{\rho_{xy} \sqrt{\rho_{XX'}} \sqrt{\rho_{YY'}}}{\sqrt{\rho_{xx'}} \sqrt{\rho_{yy'}}}$$

Lectura de la fórmula

$\rho_{XX'}$: fiabilitat millorada del test.

$\rho_{YY'}$: fiabilitat millorada del criteri.

Exemple

La correlació entre un test d'ansietat i un criteri (depressió) és de 0,63. La fiabilitat del test és de 0,79 i la del criteri de 0,90. Afegint ítems s'aconsegueix incrementar la fiabilitat del test fins a 0,83 i la del criteri fins a 0,92. Quina és l'estimació d'aquesta correlació després d'haver millorat la fiabilitat de tots dos?

$$\rho_{XY} = \frac{\rho_{xy} \sqrt{\rho_{XX'}} \sqrt{\rho_{YY'}}}{\sqrt{\rho_{xx'}} \sqrt{\rho_{yy'}}} = \frac{0,63 \sqrt{0,83} \sqrt{0,92}}{\sqrt{0,79} \sqrt{0,90}} = 0,65$$

L'estimació de la correlació entre el test i el criteri després d'haver millorat la fiabilitat de tots dos passa de 0,63 a 0,65.

7.1.5. Estimació del coeficient de validesa en el supòsit en què s'hagi millorat la fiabilitat del test

Quan només es millora la fiabilitat del test, l'estimació del nou coeficient de correlació queda determinada per:

$$\rho_{XY} = \frac{\rho_{xy} \sqrt{\rho_{XX'}}}{\sqrt{\rho_{xx'}}$$

En l'exemple anterior, si només es millora la fiabilitat del test:

$$\rho_{XY} = \frac{\rho_{xy} \sqrt{\rho_{XX'}}}{\sqrt{\rho_{xx'}}} = \frac{0,63 \sqrt{0,83}}{\sqrt{0,79}} = 0,645$$

L'estimació de la correlació entre el test i el criteri després d'haver millorat la fiabilitat del test passa de 0,63 a 0,645.

7.1.6. Estimació del coeficient de validesa en el supòsit en què s'hagi millorat la fiabilitat del criteri

L'estimació del coeficient de correlació quan només es millora la fiabilitat del criteri queda determinada per:

$$\rho_{xY} = \frac{\rho_{xy} \sqrt{\rho_{YY'}}}{\sqrt{\rho_{yy'}}}$$

En l'exemple anterior, si només es millora la fiabilitat del criteri la resposta seria:

$$\rho_{xY} = \frac{\rho_{xy} \sqrt{\rho_{YY'}}}{\sqrt{\rho_{yy'}}} = \frac{0.63 \sqrt{0.92}}{\sqrt{0.90}} = 0.637$$

L'estimació de la correlació entre el test i el criteri després d'haver millorat la fiabilitat del criteri passa de 0,63 a 0,637.

7.1.7. Valor màxim que pot assolir el coeficient de correlació entre test i criteri

Com es pot apreciar, a mesura que incrementem el coeficient de correlació del test, del criteri o de tots dos, el coeficient de correlació augmenta. Això només passa fins a cert punt, ja que el coeficient de correlació entre test i criteri sempre és més baix o igual que el seu índex de fiabilitat ($\rho_{xv} = \sqrt{\rho_{xx'}}$).

Matemàticament es pot representar de la manera següent:

$$\rho_{xy} \leq \rho_{xv}$$

Per tant, en l'exemple anterior, el màxim coeficient de correlació que es pot obtenir entre el test i el criteri és:

$$\begin{aligned} \rho_{xv} &= \sqrt{\rho_{xx'}} = \sqrt{0.79} = 0.89 \\ \rho_{xy} &\leq 0.89 \end{aligned}$$

Així doncs, el màxim valor del coeficient de correlació que es pot obtenir en aquest test és de 0,89.

7.2. Efecte de la longitud del test sobre el coeficient de correlació test-criteri

Un dels mitjans pels quals es pot incrementar el coeficient de correlació test-criteri és augmentant el nombre d'ítems que componen el test. La relació entre el nombre d'ítems i aquesta correlació és directa, és a dir, a mesura que s'incrementi el nombre d'ítems la correlació augmentarà (i es reduirà si es treuen ítems).

La relació entre tots dos queda determinada per:

$$\rho_{Xy} = \frac{\rho_{xy}\sqrt{n}}{\sqrt{1+(n-1)\rho_{xx'}}}$$

Exemple

La correlació entre un test d'ansietat de vint ítems i un criteri (depressió) és de 0,63. La fiabilitat del test és de 0,79. Estimeu el valor de la correlació test-criteri si s'afegeixen deu ítems més.

$$n = \frac{20+10}{20} = 1,5$$

$$\rho_{Xy} = \frac{\rho_{xy}\sqrt{n}}{\sqrt{1+(n-1)\rho_{xx'}}} = \frac{0,63\sqrt{1,5}}{\sqrt{1+(1,5-1)0,79}} = 0,65$$

En afegir deu ítems als vint originals, el coeficient de correlació s'incrementa des de 0,63 fins a 0,65.

Una altra possibilitat és el fet de voler arribar fins a un coeficient de correlació que es vulgui i, per tant, calgui calcular el nombre d'ítems que s'han d'afegir al qüestionari per a poder aconseguir-ho. Per a això:

$$n = \frac{(1 - \rho_{xx'})\rho_{Xy}^2}{\rho_{xy}^2 - \rho_{Xy}^2\rho_{xx'}}$$

Exemple

La correlació entre un test d'ansietat de vint ítems i un criteri (depressió) és de 0,63. La fiabilitat del test és de 0,79. Quants ítems caldrà afegir al test si es vol aconseguir un coeficient de correlació test-criteri de 0,67?

$$n = \frac{(1 - \rho_{xx'})\rho_{Xy}^2}{\rho_{xy}^2 - \rho_{Xy}^2\rho_{xx'}} = \frac{(1-0,79)0,67^2}{0,63^2 - 0,67^2 \cdot 0,79} = 2,23$$

El test s'hauria d'incrementar 2,23 vegades. Atès que el test original té vint ítems, $20 \times 2,23 = 44,6$. Evidentment, no es poden tenir decimals (estem parlant d'ítems, no es pot tenir una porció d'ítem en el test), per la qual cosa ho hem d'ajustar. Quan afegim ítems l'ajust sempre s'ha de fer cap a l'enter superior, és a dir, en aquest cas hauria d'haver-hi quaranta-cinc ítems. En cas que el test sigui excessivament llarg i no ens importi eliminar ítems fins a arribar a un coeficient de correlació test-criteri més baix (passar de 0,63 a 0,50 per exemple), l'ajust s'haurà de fer a l'enter inferior.

7.3. Efecte de la variabilitat de la mostra en la correlació test-criteri

El coeficient de correlació es veu molt afectat per la dispersió de la mostra en què estigui calculat. La relació entre la dispersió i la correlació és directa: com més dispersió s'obté més correlació.

En alguns camps de la psicologia és molt freqüent que només es pugui calcular la correlació entre el test i el criteri en una petita mostra de persones. L'exemple més clar és el de la selecció de personal. Després d'haver emprat un test per a seleccionar els candidats més adequats al lloc de treball, només

Lectura de la fórmula

ρ_{xy} : valor inicial de la correlació test-criteri.

$\rho_{xx'}$: coeficient de fiabilitat del test.

n : nombre de vegades que s'augmenta el test.

Lectura de la fórmula

ρ_{Xy}^2 : quadrat del coeficient de correlació desitjat.

es pot correlacionar la puntuació que s'ha obtingut en el test amb un criteri com el de rendiment laboral. En aquest cas, la dispersió dels seleccionats serà menor que la del total de candidats, ja que se seleccionen les persones que tenen característiques molt similars (i que s'ajusten més a les que es busquen per al lloc de treball).

Després de calcular el coeficient de correlació en la mostra de seleccionats, pot interessar estimar quin seria si s'hagués calculat sobre la totalitat d'aspirants. Per a això s'ha de partir de dos supòsits: a) el pendent de la recta de regressió és el mateix per als dos grups (admesos i aspirants) i b) l'error típic d'estimació també és el mateix per a tots dos grups. Matemàticament:

$$\begin{aligned} a) \frac{\rho_{xy}\sigma_y}{\sigma_x} &= \frac{\rho_{XY}\sigma_Y}{\sigma_X} \\ b) \sigma_y\sqrt{1-\rho_{xy}^2} &= \sigma_Y\sqrt{1-\rho_{XY}^2} \end{aligned}$$

En què les lletres majúscules fan referència al grup d'admesos, i les minúscules al total d'aspirants.

Per a estimar el valor de la correlació test-criteri en el total d'aspirants després d'haver-la calculat en el d'admesos només cal aplicar:

$$\rho_{xy} = \frac{\sigma_x \rho_{XY}}{\sqrt{\sigma_x^2 \rho_{XY}^2 + \sigma_X^2 - \sigma_X^2 \rho_{XY}^2}}$$

Exemple

Es va aplicar un test d'assertivitat a mil persones per a seleccionar deu sobrecàrrecs de vol. La desviació típica que es va obtenir en el test pel total d'aspirants va ser de 15, i en la mostra d'admesos, de 4. Després d'un quant temps treballant es va comprovar que la correlació entre les puntuacions en el test i l'execució laboral (valorada pels superiors) va ser de 0,36. Quin seria el coeficient de correlació test-criteri si s'hagués calculat sobre el total dels aspirants?

$$\rho_{xy} = \frac{\sigma_x \rho_{XY}}{\sqrt{\sigma_x^2 \rho_{XY}^2 + \sigma_X^2 - \sigma_X^2 \rho_{XY}^2}} = \frac{15 \times 0,36}{\sqrt{15^2 \times 0,36^2 + 4^2 - (4^2 \times 0,36^2)}} = 0,82$$

Com es pot veure, hi ha un increment notable en el valor del coeficient de correlació a causa de la dispersió diferent que tenen els grups.

Lectura de la fórmula

ρ_{XY} : coeficient de correlació test-criteri en la mostra d'admesos.

σ_x^2 : variància en el test del total d'aspirants.

σ_X^2 : variància en el test dels admesos.

Bibliografia

- AERA, A. N. (1974). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.
- AERA, APA, i NCME (1966). *Standards for educational and psychological test and manuals*. Washington, DC: AERA.
- AERA, APA, i NCME (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Anastasi, A. (1954). *Psychological testing*. Nova York: Macmillan.
- Beck, A., Rush, A. J., Shaw, B. F., i Emery, G. (1979). *Cognitive therapy of depression*. Nova York: Guilford Press.
- Bingham, W. V. (1937). *Aptitudes and aptitude testing*. Oxford, Anglaterra: Harpers.
- Campbell, D. T. i Fiske, A. W. (1959). Convergent and discriminant validation by the multi-trait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.
- Cronbach, L. J. (1971). Test validation. A. R. L. Thorndike. *Educational measurement* (pp. 443-507). Washington, DC: American Council on Education.
- Cronbach, L. J. i Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
- Czaja, R. i Blair, J. (1996). *Designing Surveys*. Thousand Oaks, CA: Sage Publications.
- Domènech, J. M. i Granero, R. (2003). *Anàlisi de dades per a la recerca en psicologia*. Barcelona: Signo.
- Ebel, R. (1961). Must all tests be valid? *American Psychologist*, 16, 640-647.
- Guilford, J. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 427-439.
- Hamilton, M. (1960). A rating scale for depression. *Journal Neurol. Neurosurg. Psychiatry*, 23, 56-62.
- Kane, M. (2006). Content-Related Validity Evidence in Test Development. A. S. M. Downing, i T. M. Haladyna. *Handbook of test development* (pp. 131-153). Nova Jersey: Erlbaum.
- Leighton, J. P. (2004). Avoiding misconception, missed use, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, 23, 6-15.
- Messick, S. (1989). Validity. A. R. Linn. *Educational Measurement* (3a. ed., pp. 13-104). Nova York: Macmillan.
- Muñoz, J. (2003). *Teoría clásica de los tests*. Madrid: Pirámide.
- Prieto, G. i Delgado, A. R. (2010). Fiabilidad y Validez. *Papeles del Psicólogo*, 31, 67-74.
- Ramos-Brieva, J. C. (1986). Validación de la versión castellana de la escala de Hamilton para la depresión. *Actas Luso-Esp. Neurol. Psiquiatr.*, 22, 21-28.
- Scott, W. (1917). A fourth method of checking results in vocational selection. *Journal of Applied Psychology*, 61-66.
- Shepard, L. (1993). Evaluating test validity. A. L. Darling-Hammond. *Review of research in education* (pp. 405-450). Washington, DC: American Educational Research Association.
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), 5-13.
- Shepard, L. A., Camilli, G., Linn, R., i Bohrnstedt, G. (1993). *Setting performance standards for achievement tests*. Stanford, CA: National Academy of Education.

Sireci, S. (1998). The construct of content validity. *Social Indicators Research*, 45, 83-117.

Sireci, S. G. (1998). Gathering and evaluating content validity data. *Educational Assessment*, 5(4), 299-321.

Spearman, C. (1907). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72-101.

Transformació i interpretació de les puntuacions

PID_00216189

Sergi Valero

Temps mínim de dedicació recomanat: 3 hores



**Sergi Valero**

Psicòleg del Servei de Psiquiatria de l'Hospital Universitari Vall d'Hebron de Barcelona i consultor dels Estudis de Psicologia i Ciències de l'Educació de la UOC. Doctor en Psicologia per la Universitat Autònoma de Barcelona (UAB). La seva tasca científica està emmarcada en la recerca bàsica i aplicada en l'àmbit de la psicologia, la psicopatologia i les demències. Les seves publicacions inclouen espais diversos de recerca, des de l'estudi de les dimensions de personalitat normal fins a la genètica, passant pels assajos clínics, la construcció i adaptació d'instruments de mesura psicològica, l'epidemiologia, les drogodependències o la neuropsicologia.

Índex

Introducció.....	5
1. Interpretació d'una puntuació.....	7
2. Transformació de les puntuacions.....	9
2.1. Percentils	9
2.2. Puntuacions estandarditzades	13
2.3. Puntuacions estandarditzades derivades	16
2.4. Puntuacions estandarditzades normalitzades	18
2.5. Normes cronològiques	20
3. Baremació.....	21
4. Equiparació de puntuacions.....	25
Resum.....	29
Bibliografia.....	31

Introducció

Interpretar el valor obtingut per una persona en una mesura psicològica, i fer-ho de manera raonable i raonada, no solament exigeix conèixer amb certa profunditat les característiques tècniques de l'instrument que és emprat, sinó també conèixer el constructe al qual es fa referència i entendre com aquest s'estructura i s'explica en un marc teòric que té el seu sentit, a la vegada, en un marc disciplinari concret, educacional, clínic o de qualsevol altra tipus.

És evident que un mòdul com aquest no pot abordar satisfactòriament un objectiu d'aquesta magnitud. Per aquesta raó, quan en aquest text es parli d'**interpretar una puntuació** s'haurà de considerar sempre des d'una perspectiva eminentment metodològica i matemàtica i, el més important, inferida d'acord amb la resposta d'un conjunt d'individus que, avaluats amb una mateixa mesura, proporcionaran el model de comportament que serà disposat en forma de norma, de marc de referència respecte al qual la puntuació d'una persona concreta serà interpretada. Sota aquesta perspectiva tindran el seu sentit les puntuacions transformades. Els percentils, les puntuacions estandarditzades i les normes cronològiques es recolliran aquí, exposant-ne les característiques més importants, la manera de calcular-les i, on es consideri oportú, les seves limitacions més rellevants.

Un segon terme de gran relleu, i que serà presentat en el tercer apartat, serà el concepte de *barem*. Es tracta d'una eina de gran importància en la qual s'estableix, de manera estructurada, una connexió entre la puntuació d'un individu i el comportament normatiu proporcionat per una mostra rellevant de subjectes, que, en alguns casos, apareixeran estratificats segons conveniència. Es dedicarà una atenció especial a la determinació de quina ha de ser aquesta mostra normativa. En aquest sentit s'exposaran els criteris de rellevància, representativitat i homogeneïtat com a factors fonamentals per a determinar la qualitat d'aquesta mostra de referència.

El mòdul finalitzarà exposant les diverses estratègies disponibles per a equiparar puntuacions d'instruments diferents que avaluen un mateix tret. Un conjunt d'estratègies, també metodològiques i matemàtiques, que són especialment interessants en els àmbits de l'avaluació en què els processos de memòria o aprenentatge de les persones participants poden esdevenir una limitació rellevant de les inferències que se'n deriven.

1. Interpretació d'una puntuació

Considerem una circumstància gens excepcional en què per a aprovar una assignatura, per exemple llengua anglesa, és necessari obtenir una puntuació de 5, assumint que som davant d'un examen clàssic en què la puntuació mínima possible és 0 i la màxima 10. Alguna vegada us heu preguntat quina propietat intrínseca té el número 5 per a considerar que en ell i per sobre d'ell es té assolida aquesta llengua i que, contràriament, quan s'està per sota, les nostres aptituds no satisfan el nivell exigít? Llevat que ens sentíssim confortables amb algun postulat més propi de la numerologia que no pas de la psicometria, caldria considerar que, *a priori*, el número 5 està mancat de qualsevol atribut intrínsec que el faci especial. I el mateix es podria afirmar d'una puntuació de 4 o de qualsevol altra. Si no se'ns informa de res més, haurem d'assumir que aprovar o suspendre depenent de si se supera o no una puntuació de 5 respon a una simple convenció. S'ha d'establir un punt de tall, perquè moltes vegades és necessari prendre una decisió respecte a qui supera un curs, i el valor de 5, xifra que ocupa la posició central, és el valor de referència emprat més freqüentment.

En d'altres circumstàncies, però, el criteri emprat per a discernir entre persones amb un tret determinat o sense no respon a un judici arbitrari o convencional. El criteri pot respondre a consideracions contrastades empíricament, sorgides generalment de la comparació de poblacions de subjectes ben diferenciats en el tret objecte d'interès i que permeten obtenir un punt de tall, una puntuació de referència dotada de capacitat discriminativa (Ramos-Quiroga *et al.*, 2009).

No obstant això, en moltes situacions no es disposa de criteris externs que proporcionin un marc d'interpretació respecte al qual donar rellevància o significació a una puntuació obtinguda en una mesura. De fet, es pot afirmar que la immensa majoria de les mesures psicològiques estan mancades de criteris externs i independents amb els quals inferir l'estatus actitudinal, aptitudinal o clínic d'un individu. Què es pot inferir llavors de la puntuació d'una persona en una mesura quan aquests referents no estan disponibles?

Mirar **dins** de l'instrument, és a dir, remetre's a alguna de les propietats estructurals de la mesura, és una alternativa possible. Una aproximació senzilla, per exemple, consisteix a atendre a la puntuació mínima i màxima que és possible assolir. Aquesta aproximació absoluta permet contextualitzar una puntuació tot indicant-ne l'excepcionalitat. Qualsevol manual d'ús o publicació de referència d'una mesura hauria de proporcionar informació clara respecte a aquests valors mínim i màxim. No obstant això, la senzillesa evident d'aquesta aproximació esdevé alhora la seva limitació més important: probablement tindrem menys problemes per a emetre un judici sobre una persona que obté en una mesura una puntuació de 17 que no pas sobre una persona que obté una

puntuació de 14 sabent que la puntuació màxima assolible és de 18. Ara bé, com de lluny un valor ha d'estar del valor màxim (o mínim) possible per a considerar que és molt o poc? No és possible donar una resposta operativa i universal a aquesta pregunta.

Una manera alternativa de procedir consisteix a comparar la puntuació obtinguda per una persona amb la puntuació que obtenen altres persones que han estat avaluades amb la mateixa mesura. La mitjana (i també la mediana) és una manera freqüent de resumir el comportament d'aquest grup de persones. D'aquesta manera, si una persona obté en una mesura d'aptituds mnèsiques una puntuació de 7 i la mitjana dels seus col·legues va ser de 9, avaluats seguint la mateixa pauta, podem afirmar que aquesta persona presenta una capacitat de memòria inferior a la que presenten els seus companys. I aquí descansa una de les claus d'aquesta aproximació relativa: la inferència que es fa d'una observació concreta és una funció de la tendència de comportament d'altres persones en la mateixa mesura. Reprendrem aquesta observació més endavant.

2. Transformació de les puntuacions

D'acord amb el que s'ha dit fins ara, és possible disposar d'estratègies que permeten la interpretació d'una mesura i que, a més, no impliquen cap manipulació de les dades obtingudes. No obstant això, no són les úniques estratègies a l'abast per a donar sentit a una puntuació determinada. Altres procediments, menys simples que els exposats fins ara, ofereixen a l'usuari una aproximació més precisa i informativa. No obstant això, aquests procediments impliquen la transformació de les puntuacions observades.

Transformar les puntuacions d'una mesura consisteix a aplicar una estratègia de codificació en què les puntuacions obtingudes en la mesura, les seves puntuacions directes, són recodificades (transformades) en un nou sistema de valors que en faciliten a l'usuari la interpretació.

Una propietat necessària de les puntuacions transformades és que no alteren l'escalament de les puntuacions directes, és a dir, respecten la diferent disposició dels diferents subjectes d'acord amb la magnitud de les seves puntuacions. I aquest manteniment de l'ordenació original es porta a terme, com s'ha esmentat, atorgant a les noves puntuacions un sentit eminentment pràctic. Els percentils, les puntuacions estandarditzades, les puntuacions estandarditzades derivades i normalitzades i les normes cronològiques seran les diverses estratègies de transformació que es tractaran a continuació.

2.1. Percentils

Per a descriure el concepte de *percentil*, també conegut com a *centil*, s'exposarà en primer lloc el concepte de *mediana*. Un cop ordenats els valors d'un conjunt d'observacions, la mediana serà el valor que ocupa la posició o ordre central. Un exemple senzill ajudarà en aquesta primera aproximació.

Exemple

Es disposa del conjunt de valors següent: 4, 5, 4, 3, 7, 8, 3, 1 i 6. En primer lloc, caldrà ordenar la seqüència: 1, 3, 3, 4, 4, 5, 6, 7, 8. La mediana d'aquest conjunt de valors és el 4 (el segon 4), ja que és el valor que ocupa la posició central de la distribució. La meitat de les persones són disposades a un costat i l'altre d'aquest valor.

En aquest exemple ha estat fàcil identificar el valor central de la distribució atès que el nombre d'observacions és senar. Tota distribució imparella té un valor que ocupa la posició central. I si la distribució de valors fos parella? Afegim a la distribució de valors ja coneguda i ordenada un nou valor, qualsevol: 1, 3, 3, 4, 4, 5, 6, 7, 8 i 9. En una distribució amb un nombre parell d'observacions no són un, sinó dos, els números que s'ubiquen en la posició central. En el cas que ens ocupa són els valors 4 i 5. Aplicant una mitjana aritmètica entre els

dos valors es conclourà que la mediana d'aquesta nova distribució de valors és 4,5. Tot és correcte: una mediana que prové de valors enters pot ser decimal i, a més, no ha coincidir necessàriament amb cap de les observacions originals.

No obstant això, què tenen a veure els percentils amb la mediana? Els percentils són una generalització de la mediana. La transformació basada en els percentils consisteix a assignar a cada puntuació directa una puntuació percentual d'acord amb la posició de les observacions dins del conjunt d'observacions. Si el valor percentual és del 25% s'estarà parlant del percentil 25. I sota aquesta denominació s'estarà identificant la puntuació directa que deixa per sota una quarta part de totes les observacions. O expressat d'una altra manera, el valor que amb ell i per sobre d'ell hi ha el 75% de totes les observacions. Si el percentatge ascendeix a un 50% es parlarà llavors del percentil 50, que es correspondrà amb el valor que ocuparà la posició central de la distribució. El percentil 50 és, efectivament, la mediana.

La primera formulació pràctica dels percentils es remunta al final del segle XIX i prové de l'explorador i científic Francis Galton, cosí de Charles Darwin que, entre altres aficions, es va mostrar força interessat per la mesura de multitud de trets antropomètrics. Avui en dia, i sense abandonar l'àmbit de la mesura de les propietats del cos, els percentils continuen essent una estratègia d'interès, per exemple, per als pares i mares que volen saber en quin grau els seus fills s'ajusten als altres nens d'edat comparable en termes d'alçada i pes.

Els percentils han esdevingut una mesura molt emprada no solament en el context de la mesura del pes o de l'alçada, sinó també en el context de les mesures psicològiques. La seva simplicitat i universalitat, ja que no cal tenir coneixement amplis de matemàtiques ni d'estadística en particular, n'ha estimulat sens dubte la difusió com a estratègia interpretativa. Exposem un cas pràctic que servirà per a exemplificar-ne el sentit i generalitzar-ne el càlcul.

En una mostra de 1.000 persones de nivell acadèmic, edat i gènere diferents (Gomà-i-Freixament *et al.*, 2008) es va administrar el *Zuckerman-Kuhlman Personality Questionnaire* (ZKPQ, Zuckerman, Kuhlman, Joireman, Teta i Kraft, 1993), un instrument que mesura cinc dimensions bàsiques de la personalitat normal.

Entre les cinc dimensions que el model permet mesurar, s'ha centrat l'atenció, per motius de simplificació, en una de les variables de personalitat, el neuroticisme¹. En aquest instrument, aquesta dimensió de personalitat és operativitzada mitjançant un total de 19 ítems de resposta tipus veritable-fals, en què el sumatori de tots (tenint en compte que alguns ítems són directes i altres inversos) esdevé una estimació del grau en què un individu tendeix a presentar aquest tret de personalitat. Cal estar atent a les peculiaritats de cada instrument per a saber com s'ha d'obtenir una puntuació en la variable d'interès. No sempre se segueix una mateixa pauta per a calcular les puntuacions totals.

En el cas que ens ocupa, la puntuació màxima assolible és 19, i la mínima 0. Una puntuació més elevada equival a una intensitat més elevada del tret. La distribució de valors absoluts i relatius d'aquest grup de 1.000 persones en aquesta variable apareix recollida en la taula 1.

⁽¹⁾El neuroticisme expressa la tendència de tota persona a mostrar-se alterada, amb tensió, preocupada, amb indecisió i amb labilitat emocional (estat d'humor freqüentment canviant).

Taula 1. Puntuacions directes, freqüències, percentatges i percentils

Puntuació en l'escala	Freqüència absoluta (fi)	Freqüència acumulada (fa)	Percentatge (%)	Percentatge acumulat (%)	Percentil (Pc)
0	21	21	2,1	2,1	2
1	55	76	5,5	7,6	3
2	49	125	4,9	12,5	8
3	77	202	7,7	20,2	13
4	63	265	6,3	26,5	21
5	70	335	7,0	33,5	27
6	76	411	7,6	41,1	34
7	67	478	6,7	47,8	42
8	68	546	6,8	54,6	48
9	72	618	7,2	61,8	55
10	59	677	5,9	67,7	62
11	60	737	6,0	73,7	68
12	57	794	5,7	79,4	74
13	44	838	4,4	83,8	80
14	43	881	4,3	88,1	84
15	41	922	4,1	92,2	89
16	38	960	3,8	96,0	93
17	23	983	2,3	98,3	97
18	11	994	1,1	99,4	99
19	6	1000	0,6	100,0	99

Si es pren per exemple la puntuació de 8, es pot observar que han estat 68 les persones que han presentat aquesta puntuació (columna de freqüència absoluta). Atès que són un total de 1.000 persones en aquesta mostra, es pot afirmar que un 6,8% dels participants tenen una puntuació directa de 8 (columna de percentatge). La penúltima columna, identificada com a percentatge acumulat, ha exigut que els valors directes fossin disposats de manera ordenada. Els percentatges corresponents a cada puntuació es van sumant i acumulant. En el cas de la puntuació directa 8, el percentatge acumulat és de 54,6. Això vol dir que el 54,6% de les persones de la mostra ha obtingut puntuacions de com màxim 8 unitats (8 o menys). Quin és el percentil associat a aquest valor directe? Atès que ha de ser el percentatge de persones que queden per sota d'aquest valor, caldrà observar que fins a arribar a una puntuació directa de 7 hi ha un percentatge acumulat del 47,8%. Aquest valor, un cop arrodonit a l'enter immediatament superior, passa a ser d'un 48%. La puntuació directa 8 s'associa a un percentil de 48.

Si la correspondència entre puntuacions directes i percentils s'enfoqués de manera inversa, en canvi, es podria formular una pregunta del tipus "a quin valor directe correspondria un percentil de 55?" Segons percentatges acumulats, fins a un valor directe de 8 hi correspon un percentatge acumulat de 54,6%. Un percentatge de 55, tot i que per poc, encara no s'ha produït. El valor directe següent és el 9, i fins a aquest s'ha acumulat el 61,8% de les observacions. En aquest cas sí, el 55% objecte

d'interès ja s'ha produït, ja que està inclòs en el 61,8%. Per tant, la puntuació directa 9 s'associa a un percentil de 55.

D'acord amb aquest raonament caldrà entendre per què una puntuació de 9 és també percentil 56. Però també 57, 58, 59, 60 i 61. Només quan s'arriba a la puntuació directa 10 el percentil passa a ser de 62. Aquesta circumstància, en què un mateix valor directe es vincula a més d'un percentil, és del tot plausible quan la variable objecte d'interès no disposa de prou valors per a cobrir les cent unitats per a les quals el percentil sí està preparat.

Cal tenir en compte que aquesta estratègia per a calcular percentils, basada en els percentatges acumulats, no és l'única possible. Hi ha altres aproximacions que també són acceptades. I diem aproximacions perquè l'estudiant haurà de tolerar que no totes arribin necessàriament a un resultat idèntic. Una estratègia alternativa a la ja exposada per a calcular un percentil, qualsevol, és la que es proposa a continuació:

$$P = \frac{fa + 0,5fi}{N} \times 100$$

Així doncs, i d'acord amb la taula 1, el percentil que s'associa a una puntuació directa de 9 és 58:

$$P_9 = \frac{546 + (0,5 \times 72)}{1000} \times 100 \approx 58$$

Atès que els percentils no tenen decimals, s'ha ignorat el decimal del resultat exacte obtingut (58,2). Cal recordar que, seguint la primera estratègia exposada per a calcular percentils, la puntuació directa 9 era, alhora, els percentils 55, 56, 57, 58, 59, 60 i 61. El valor obtingut en la fórmula anterior és només un dels valors, el central.

El primer dels avantatges de l'ús dels percentils, com ja hem indicat, és la simplicitat conceptual. Un segon avantatge és que per a calcular-lo i interpretar-lo correctament, atès que centra l'atenció només sobre l'ordre en què les observacions són disposades, esdevé irrellevant la manera com es distribueixen els valors objecte d'anàlisi. Això implica que, tot i que hi ha estratègies de transformació que requereixen una distribució de valors essencialment ajustada a una corba normal, més endavant es veurà, en el cas dels percentils és possible abordar qualsevol tipus de distribució. Aquesta característica és especialment interessant en psicologia, en què moltes de les variables d'interès rarament s'ajusten a una distribució normal de valors.

Com s'ha vist, els percentils consisteixen a dividir una distribució de valors en 100 posicions ordinals. No obstant això, hi ha altres abordatges que simplifiquen la quantitat d'ordres possibles i que poden ser suficients per a resumir el comportament d'un conjunt de valors. Els més comuns són:

Lectura de la fórmula

f_a : freqüència acumulada prèvia a la puntuació directa de la qual es vol calcular el percentil.
 f_i : freqüència absoluta en la qual es troba la puntuació directa.
 N : nombre de persones que constitueixen la mostra.

Enllaç recomanat

En la Viquipèdia (en anglès) i en d'altres recursos a Internet, trobareu estratègies alternatives, també vàlides, per a calcular els percentils.

- Els **quartils**. Divideixen una distribució en quatre parts. Se solen identificar com a *quartil 1*, *quartil 2* i *quartil 3*. Coincideixen amb els percentils 25, 50 i 75, respectivament.
- Els **quintils**, que dividint la distribució en cinc parts, s'equiparen als percentils 20, 40, 60 i 80.
- Els **decils**. Representen una divisió en deu parts i es corresponen amb els percentils 10, 20, 30, 40, etc.

2.2. Puntuacions estandarditzades

Si bé la mitjana és un estadístic de tendència central que permet expressar el comportament mitjà d'un conjunt d'observacions, la desviació típica (o estàndard) permet quantificar com de diferents són aquestes observacions. Aquest índex de dispersió expressa la discrepància mitjana entre una observació qualsevol i la mitjana del conjunt d'observacions. La mitjana i la desviació típica esdevenen la matèria primera imprescindible per a calcular una puntuació estandarditzada. Si bé els percentils dirigeixen l'atenció cap a la posició que ocupa un individu respecte a un grup de referència, en el cas de les puntuacions estandarditzades l'interès es focalitza en la discrepància que presenta aquest individu respecte a la forma de comportament mitjà del grup, la seva mitjana.

La puntuació estandarditzada o puntuació típica, és la que permet expressar quantes desviacions típiques una observació se situa per sobre o per sota d'una mitjana. Aquest procés d'estandardització té el seu sentit quan la distribució de valors objecte d'interès s'ajusta essencialment a una corba normal. El valor resultant d'aquesta transformació es constitueix en una nova puntuació que rep el nom de *puntuació estandarditzada z*. El seu càlcul respon a la fórmula següent:

$$Z_x = \frac{X - \bar{X}}{S_x}$$

Cal tenir present que aquesta puntuació *z* tindrà sempre una mitjana de 0 i una desviació estàndard d'1. En la taula 2, i reprenent l'estudi sobre neuroticisme en una mostra de 1.000 persones:

Lectura de la fórmula

X : puntuació directa.
 $\bar{X}$: mitjana de la mostra.
 S_x : desviació típica de la mostra.

Taula 2. Puntuacions directes, freqüències i puntuacions estandarditzades *Z*

Puntuacions directes	Freqüència absoluta	Puntuacions estandarditzades <i>z</i>
0	21	-1,72
1	55	-1,51
2	49	-1,30
3	77	-1,09

Puntuacions directes	Freqüència absoluta	Puntuacions estandarditzades z
4	63	-0,88
5	70	-0,66
6	76	-0,45
7	67	-0,24
8	68	-0,03
9	72	0,18
10	59	0,39
11	60	0,61
12	57	0,81
13	44	1,03
14	43	1,24
15	41	1,45
16	38	1,67
17	23	1,88
18	11	2,09
19	6	2,30

Per a poder calcular els valors z, ubicats en la darrera columna de la taula, és necessari saber la mitjana i la desviació típica de la variable en aquesta mostra: la mitjana era 8,14 i la desviació típica 4,72. Així doncs, en el cas concret d'una persona que ha obtingut una puntuació directa de 8, la seva puntuació z serà:

$$-0,03 = \frac{8 - 8,14}{4,72}$$

Què significa aquest valor de -0,03? Atès que la mitjana de z sempre és 0 i que la seva desviació típica és també sempre 1, observar una puntuació típica de -0,03 significa que els subjectes que tenen una puntuació directa de 8, i que són un total de 68 persones, s'ubiquen a 0,03 desviacions típiques de la mitjana de la mostra. Sabem, a més, que aquest valor directe s'ubica per sota de la mitjana perquè la seva puntuació típica és negativa. Necessàriament, tots els subjectes que presenten puntuacions directes que es disposen per sota de la mitjana obtindran puntuacions z negatives, mentre que els que s'ubiquen per sobre de la mitjana presenten puntuacions típiques positives. Si ens fixem en la part final de la taula 2, veurem que les 6 persones que obtenen una puntuació de 19 tenen una puntuació típica de 2,30, la qual cosa implica que aquestes s'ubiquen a més de 2 desviacions típiques per sobre de la mitjana. En canvi, les

21 persones que obtingueren una puntuació directa de 0 presenten una puntuació típica de $-1,72$. Són persones que queden, per tant, a $1,72$ desviacions per sota de la mitjana.

Una pregunta important que ens podem formular és si distanciar-se en $1,72$ o $2,30$ desviacions típiques d'una mitjana és molt o poc. És clar que allunyar-se d'una mitjana en $2,30$ desviacions típiques és allunyar-se més que $1,72$ vegades. En la taula 2 podeu veure que en aquesta mostra les persones amb la màxima puntuació es disposen més allunyades de la mitjana que les persones que obtenen la puntuació més baixa possible. Ara bé, és molt fer-ho $2,3$ vegades?

En l'estudi de Valero *et al.* (2012) dues mostres de subjectes adults són comparades. Una constituïda per 217 persones adultes afectades per un trastorn per dèficit d'atenció amb hiperactivitat (TDAH), i l'altra per 434 subjectes control aparellats per edat i gènere. Es van contrastar diverses variables de personalitat, però per simplificar l'exposició se centrà de nou l'atenció sobre la dimensió de neuroticisme. La taula 3 exposa les puntuacions típiques en aquesta variable per a cadascun dels dos grups d'estudi.

Taula 3. Puntuacions directes i estandarditzades per a una mostra de persones amb TDAH i controls

Puntuacions directes	Puntuació estandarditzada z	
	TDAH	Controls
0	-2,54	-1,67
1	-2,38	-1,46
2	-2,16	-1,25
3	-1,94	-1,04
4	-1,73	-0,83
5	-1,51	-0,62
6	-1,30	-0,41
7	-1,08	-0,19
8	-0,87	0,02
9	-0,65	0,23
10	-0,43	0,44
11	-0,22	0,65
12	0	0,86
13	0,21	1,08
14	0,43	1,29
15	0,65	1,50
16	0,86	1,71

Puntuacions directes	Puntuació estandarditzada z	
17	1,08	1,92
18	1,29	2,14
19	1,51	2,35
	Mitjana = 12,01	Mitjana = 7,91
	Desviació típica = 4,63	Desviació típica = 4,72

Com es pot veure a peu de taula, els subjectes afectats pel TDAH presenten una mitjana en la mesura de neuroticisme superior a la presentada pels subjectes control (12,01 enfront de 7,91). Es tracta d'un resultat esperable tenint en compte que una és mostra psiquiàtrica i l'altra no. Les desviacions típiques respectives, en canvi, i en aquest estudi, són força comparables. Si per a cada puntuació directa es comparen les puntuacions típiques respectives de les dues mostres es veurà que no coincideixen en cap cas. De fet, es mostren clarament diferenciades.

En el cas de les persones afectades pel TDAH les puntuacions tipificades tendeixen a estar sistemàticament desplaçades cap a la part baixa de la distribució respecte a les presentades pels subjectes control. La discrepància entre les dues mitjanes condiciona aquest comportament. Si bé una puntuació directa de 0 en el cas dels subjectes control implica ubicar-se a 1,67 desviacions típiques per sota de la mitjana, en el cas dels pacients afectats de TDAH, la mateixa puntuació directa els allunya més de 2 desviacions típiques i mitja de la mitjana respectiva. Tot depèn del grup en què s'estigui ubicat.

I aquí descansa una de les consideracions més rellevants que cal tenir presents quan s'interpreten puntuacions estandarditzades: les inferències a les quals podem arribar mitjançant l'ús d'aquestes puntuacions depenen de la mostra que és emprada, ja que és d'aquesta que s'obtenen la mitjana i la desviació típiques que permeten el càlcul de les puntuacions transformades. Triar acuradament la mostra de referència esdevé imprescindible. Més endavant, quan es parli del concepte de *barems*, es reprendrà aquesta idea.

2.3. Puntuacions estandarditzades derivades

Encara que intrínsecament no és un problema, les puntuacions z s'expressen amb valors negatius i decimals. El sentit bàsic de les puntuacions estandarditzades derivades és transformar les puntuacions típiques d'acord amb un sistema de codificació que elimina aquests valors negatius i decimals. Una de les puntuacions típiques derivades més comunament emprades és la puntuació *T* de McCall. La seva formulació és senzilla:

$$T = 50 + 10z$$

Consisteix a partir d'una constant, en aquest cas 50, i sumar-hi deu vegades el valor de la puntuació típica z . Per a interpretar degudament aquesta puntuació només cal tenir present que la mitjana d'aquesta puntuació és 50 i la seva desviació típica 10. D'aquesta manera, per exemple, una puntuació z de 2 tindrà un valor de 70 un cop transformada en puntuació T . En tots dos casos (z i T), i necessàriament, seria indicador d'una puntuació directa que s'ubica per sobre de la mitjana corresponent en 2 desviacions típiques.

La taula 4 recull els valors de la taula 2 afegint les puntuacions T corresponents:

Taula 4. Puntuacions directes, freqüències i puntuacions estandarditzades z i T

Puntuacions en l'escala	Freqüència absoluta	Puntuació estandarditzada z	Puntuació estandarditzada derivada T
0	21	-1,72	33
1	55	-1,51	35
2	49	-1,30	37
3	77	-1,09	39
4	63	-0,88	41
5	70	-0,66	43
6	76	-0,45	46
7	67	-0,24	48
8	68	-0,03	50
9	72	0,18	52
10	59	0,39	54
11	60	0,61	56
12	57	0,81	58
13	44	1,03	60
14	43	1,24	62
15	41	1,45	65
16	38	1,67	67
17	23	1,88	69
18	11	2,09	71
19	6	2,30	73

És possible trobar puntuacions típiques derivades de naturalesa diversa, i és que, estrictament, tothom se'n pot crear una de pròpia, ja que una puntuació derivada és formulada d'acord amb criteris arbitraris. Exemples diferents els trobarem en el test d'intel·ligència Stanford-Binet, que empra una mitjana de 100 i una desviació típica de 16; la *Wechsler Adult Intelligence Scale* (WAIS), amb

una mitjana també de 100, però amb una desviació típica de 15; o el *Minnesota Multiphasic Personality Inventory* (MMPI), que, coincidint amb l'escala de McCall, s'interpreta entorn a una mitjana de 50 i una desviació típica de 10.

2.4. Puntuacions estandarditzades normalitzades

Si s'assumeix que la distribució de valors que és objecte d'anàlisi s'ajusta a una corba normal, és possible transformar els percentils en les puntuacions estandarditzades que es correspondrien si la distribució fos, efectivament, normal. Aquestes puntuacions rebrien el nom de *puntuacions estandarditzades normalitzades*.

Taula 5. Puntuacions directes, freqüències, puntuacions estandarditzades z i T i normalitzades

Puntuació en l'escala	Freqüència absoluta (f_i)	Freqüència acumulada (F_a)	Percentatge acumulat	Percentil (P_c)	Puntuació estandarditzada z	Puntuació estandarditzada normalitzada
0	21	21	2,1	2	-1,72	-2,05
1	55	76	7,6	3	-1,51	-1,88
2	49	125	12,5	8	-1,30	-1,41
3	77	202	20,2	13	-1,09	-1,13
4	63	265	26,5	21	-0,88	-0,81
5	70	335	33,5	27	-0,66	-0,61
6	76	411	41,1	34	-0,45	-0,41
7	67	478	47,8	42	-0,24	-0,20
8	68	546	54,6	48	-0,03	-0,05
9	72	618	61,8	55	0,18	0,13
10	59	677	67,7	62	0,39	0,31
11	60	737	73,7	68	0,61	0,47
12	57	794	79,4	74	0,81	0,64
13	44	838	83,8	80	1,03	0,84
14	43	881	88,1	84	1,24	0,99
15	41	922	92,2	89	1,45	1,23
16	38	960	96,0	93	1,67	1,48
17	23	983	98,3	97	1,88	1,88
18	11	994	99,4	99	2,09	2,33
19	6	1.000	100,0	99	2,30	—

D'acord amb la taula 5, si es pren per exemple la puntuació directa 9 s'observarà que, com ja se sabia, la seva puntuació estandarditzada z és de 0,18, i el seu percentil de 55. El pas següent és cercar una taula estandarditzada de

⁽²⁾En un cercador teclegeu, per exemple, z tables.

puntuacions normals, també coneguda com a *taula de puntuacions z*. A Internet en trobareu moltes². La taula 6 és l'extracte d'una d'aquestes taules. Si s'hi cerca una proporció de 0,55 es localitzarà el valor z corresponent.

En el cas que ens ocupa, la proporció més propera és 0,5517, i en conseqüència el valor z que s'hi vincula és 0,13. Qualsevol percentil, en forma de proporció, té associat un valor z i viceversa.

Taula 6. Extracte taula puntuacions z

z	0,0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224

Reiterem que el valor obtingut en una taula de valors z assumeix que la distribució de valors que és objecte d'interès s'ajusta bàsicament a una corba normal. La discrepància observada entre les puntuacions 0,18 i 0,13 (vegeu la taula 5), o genèricament entre qualsevol puntuació estandarditzada calculada i la corresponent puntuació estandarditzada normalitzada, respon a la discrepància entre la distribució de valors empírics i una distribució teòrica normal. Cal tenir present que la discrepància entre les dues distribucions de valors tipificats no serà l'excepció, sinó la norma. Com ja s'ha esmentat amb anterioritat, són moltes les variables que no s'ajustaran a una (perfecta) distribució normal. Per a les que ho facin essencialment, heus aquí una manera possible de transformar una puntuació.

De la mateixa manera que en el cas de les puntuacions estandarditzades derivades és possible disposar també de puntuacions estandarditzades normalitzades derivades. En psicologia, les més habituals són les estannines o enneatipus i els decatipus. La formulacions respectives són:

$$\text{Enneatipus} = 5 + 2z_n$$

$$\text{Decatipus} = 5,5 + 2z_n$$

En què la z és la puntuació estandarditzada normalitzada.

Assumint una distribució normal de valors, si cliqueu en un cercador termes com *z distribution percentile* podreu veure de manera gràfica la correspondència que s'estableix entre alguns dels procediments de transformació presentats en aquest mòdul.

2.5. Normes cronològiques

Una altra possible transformació de les puntuacions directes consisteix a fer-ho en forma de normes cronològiques o d'edats. L'estratègia a seguir consisteix a administrar a diversos grups d'edat, representatius d'una determinada població, una mateixa mesura, recollint per a cada grup d'edat la mitjana respectiva. Posteriorment, quan un individu és avaluat, la seva puntuació és transformada en una edat, l'edat que en la població de referència s'aproximaria a la puntuació obtinguda. La idea bàsica que justifica aquesta fórmula és simple: assumir que hi ha formes de comportament que tendeixen a ser típiques entorn d'una edat. Així doncs, si després de seleccionar una mostra representativa d'una determinada població, s'administra una determinada mesura als nens de sis anys i la mitjana obtinguda és de 12, i la mateixa mesura administrada als nens de vuit anys és de 17, si un nen concret obté una puntuació de 17 s'afirmarà que en aquesta mesura el nen té una edat de vuit anys (encara que cronològicament en tingui set).

Un dels contextos en què aquesta estratègia ha esdevingut més popular ha estat en parlar del quocient d'intel·ligència. La seva formulació és:

$$QI = \frac{EM}{EC} \times 100$$

Lectura de la fórmula

EM: edat mental.
EC: edat cronològica.

La transformació en edats és una estratègia antiga, probablement tant com ho ha estat la mesura dels constructes psicològics. No obstant això, no es pot deixar d'esmentar que aquesta transformació s'acompanya de limitacions que no es poden menystenir i que al llarg del temps n'han anat reduint la presència en la bibliografia. Una de les seves limitacions més rellevants és que el model assumeix que els canvis que es van produint al llarg del temps, segons s'analitzen diversos grups d'edat, són canvis essencialment lineals, és a dir, que el canvi produït entre les persones de sis a set anys és comparable als canvis que es produeixen entre les persones de vuit respecte les de nou anys. Malauradament, assumir aquest principi és excessiu per a molts processos psicològics. Canvis produïts entre determinades edats cronològiques no tenen necessàriament la mateixa rellevància o significat al llarg de totes les edats mentals ni, per descomptat, a l'inrevés.

Lectura recomanada

A Fancher (1987) trobareu un text interessant en el qual s'aborda no solament el concepte d'edat mental, sinó que també es fa un repàs històric del concepte d'intel·ligència i també d'algunes controvèrsies entorn de la seva mesura.

3. Baremació

La baremació, també coneguda com a *escalament d'una mesura*, és un element de gran rellevància en el procés d'elaboració i ús d'un instrument de mesura. Com s'ha vist en aquest mòdul, la transformació d'una puntuació ajuda l'usuari a injectar significat a un determinat valor d'acord amb el comportament d'un conjunt d'observacions. Si l'instrument disposa d'un manual, aquest hauria de reportar els barems de la mesura, amb les transformacions necessàries, si s'escau, a més de les indicacions bàsiques per a interpretar degudament qualsevol puntuació. En d'altres casos no hi haurà manual de referència, però sí probablement una o més publicacions periòdiques a les quals ens podrem remetre i trobar la informació necessària.

Independentment del tipus de publicació amb què el barem d'una mesura es reporti, perquè tota interpretació que se'n derivi es faci segons un estàndard de qualitat elevat, cal que la mostra que proporciona les dades que permeten satisfer una transformació (per exemple, mitjanes i desviacions típiques) i que rebrà el nom de *mostra normativa*, sigui una mostra rellevant, representativa i homogènia.

Rellevant perquè cal garantir que la mostra que conforma els barems d'un instrument fa referència a una població que és significativa d'acord amb la satisfacció d'uns objectius específics. Tal com s'ha vist durant l'exposició de les puntuacions estandarditzades, el sentit que pot tenir una determinada puntuació tipificada ha de ser entès sempre dins del marc de la població que proporciona els paràmetres de referència. No hi ha una població *a priori* ideal sobre la qual es pugui desenvolupar el barem d'un instrument de mesura. La població ideal depèn del tipus de persona que es vol descriure i, sobretot, dels termes en què es proposi fer-ho.

Probablement s'estaria d'acord que si es vol inferir alguna cosa rellevant sobre el grau de coneixement dels joves d'un institut sobre la Revolució Industrial seria convenient poder disposar d'una mostra àmplia d'altres joves provinents d'altres instituts amb els quals contrastar-ne el rendiment. No obstant això, no sempre és possible identificar una única població respecte a la qual es pugui inferir satisfactòriament l'estat d'un individu.

D'acord amb la mateixa línia argumental exposada en parlar del neuroticisme i TDAH, si es vol interpretar la intensitat de la simptomatologia depressiva d'una persona mitjançant un determinat instrument clínic, una puntuació concreta no té la mateixa consideració si els barems emprats provenen de població general, que si provenen de diversos centres d'atenció primària o

d'unitats d'hospitalització psiquiàtrica. *A priori* les tres poblacions són igualment vàlides, en tant que en qualsevol d'elles pot ser pertinent per a obtenir dades normatives i interpretar la gravetat de la simptomatologia depressiva d'aquesta persona. Ara bé, no és el mateix ubicar-se en un percentil de 70 prenent com a mostra normativa població general que població clínica. L'excepcionalitat (o no) de la simptomatologia d'aquest individu dependrà necessàriament de quina sigui la població de referència que s'expressa mitjançant els barems, i això condicionarà la interpretació que se'n pugui derivar i les decisions que caldria prendre respecte a aquesta (McIntire i Miller, 2007). Cal delimitar degudament, en forma d'objectius de treball clars, què ha de ser avaluat i, sobretot, amb quina finalitat.

S'ha afirmat també que la mostra constitutiva dels barems ha de ser **representativa**. Un cop definida quina ha de ser la població diana, l'estratègia emprada per a seleccionar les persones de la població que formaran part de la mostra normativa esdevé ara el focus d'atenció principal. D'acord amb els estàndards que proporciona l'American Psychological Association (APA, 1999), cal exposar amb cura i claredat quina ha estat la població objectiu i, alhora, la mostra que se n'ha seleccionat. Només així és possible delimitar el tipus d'inferència a la qual es pot arribar quan s'empren uns determinats barems.

Un mostreig probabilístic és la millor estratègia metodològica disponible per a assegurar l'obtenció d'una mostra representativa, ja que només és en aquest context aleatori en el qual qualsevol individu de la població disposa d'una probabilitat idèntica de formar part de la mostra. I tal com exposa Aiken (1996), el mostreig aleatori estratificat hauria de ser l'estratègia d'elecció. Aquest tipus de mostreig hauria de proporcionar la deguda representativitat per a tots els paràmetres que han estat considerats rellevants. En aquest sentit, l'edat i el gènere són alguns dels factors que més freqüentment articulen un barem, però qualsevol altre variable pot ser objecte d'interès i per tant d'estratificació. L'àmbit d'estudi i els objectius que han de ser resolts condicionaran el tipus de variables que se sotmeten a escalament i el grau d'estratificació amb el qual aquestes variables s'haurien d'expressar.

Malauradament no sempre s'explicita en els textos especialitzats, però no es pot deixar d'esmentar que no sol ser fàcil satisfer un mostreig estrictament probabilístic. Són moltes les condicions que, en context aplicat, minven l'èxit d'un mostreig aleatori. Considerem per exemple la presència de persones institucionalitzades a les quals difícilment es tindrà accés (presos, persones hospitalitzades, etc.), les que no tenen les aptituds necessàries per a donar resposta a una determinada mesura (analfabetes, amb problemes de comprensió, etc.), les avaluades en contextos en què hi pot haver una alt desig social (selecció de personal, presència de preguntes compromeses, etc.), persones que simplement no volen col·laborar, etc. Els barems de la majoria d'instruments d'avaluació, si no tots, estan subjectes a circumstàncies que limiten la generalització dels seus valors. L'esforç de tot investigador hauria de dirigir-se a mi-

nimitzar o atenuar en la mesura del possible aquests biaixos i, en tot cas, a informar sempre detingudament els usuaris de l'instrument sobre què s'ha fet i com s'ha resolt.

Quan es parla de l'**homogeneïtat** de la mostra es fa referència a la necessitat de controlar la presència d'efectes de confusió o estranys que puguin esbiaixar o fins i tot invalidar les interpretacions que se'n derivin. Les persones que seran avaluades mitjançant una determinada mesura haurien de ser comparables a les persones que van formar part de la mostra de participants que en van constituir el barem. Discrepàncies consistents entre les dues mostres poden derivar en inferències que sobreestimarien o infraestimarien l'estat real de les persones avaluades.

Aquestes mateixes consideracions justifiquen la necessitat que hi hagi, també, una actualització contínua de les dades que estructuren els barems. Caldria poder disposar de barems que se sotmetin a revisions amb una determinada regularitat. I és que caldria assumir per defecte que les dades contingudes en un barem caduquen. Les variables que són objecte d'interès són dinàmiques i, per tant, canviants, modulades pel decurs dels temps. Són sensibles a les transformacions científiques, socials i culturals que es van esdevenint i, per aquesta raó, tot barem hauria de poder reflectir de manera fidel l'estat actual del concepte que és objecte de consideració. Tot i que les consideracions teòriques i empíriques són diferents, emprar actualment barems de qualitat de vida en mostres i sota criteris de fa trenta anys, per exemple, seria comparable a emprar barems d'alçada incloent-hi persones avaluades durant els anys seixanta. En qualsevol dels dos casos s'estaria cometent un error o, en el millor dels casos, un biaix difícilment assumible.

L'adaptació de l'instrument a les característiques lingüístiques i culturals de les persones que han de ser avaluades ha de ser també motiu d'una necessària atenció. En la majoria de casos, els instruments d'avaluació que són emprats en psicologia en el nostre entorn cultural tenen l'origen en el món anglosaxó. L'avantatge que presenta aquesta circumstància és que permet que, un cop traduït un instrument, esdevingui més fàcil l'homologació de l'instrument des d'un punt de vista internacional, alhora que facilita la satisfacció d'estudis de tipus transcultural en els quals un mateix tret es contrasta entre diversos grups culturalment diferents. No obstant això, cal tenir present que traduir un instrument és condició necessària però no sempre suficient. Seguint les indicacions dels estàndards de l'American Psychological Association (APA, 1999), cal que hi hagi una adaptació cultural adequada que sigui sensible a les peculiaritats dels individus que seran avaluats amb l'objectiu. Una adaptació correcta de la mesura constitueix un objectiu dirigit, de nou, a proporcionar estimacions sense biaix i que, per tant, permeti inferir amb garanties suficients de validesa i fiabilitat l'estat dels individus.

En la configuració del barem d'un instrument és possible emprar qualsevol de les estratègies exposades en aquest mòdul. Les dades descriptives, disposades en taules, es poden expressar de manera transformada, per exemple mitjançant percentils, i agrupades segons estratificacions per edat (Rodríguez *et al.*, 2012). En d'altres casos, els percentils es disposaran sense estratificació (Alegret *et al.*, 2012; Romero *et al.*, 2000). En algunes publicacions s'oferiran mitjanes i desviacions típiques que permetran, si es vol, una transformació en puntuacions estandarditzades. I ho poden fer d'acord amb dos factors, per exemple, l'edat i el gènere (Gomà-i-Freixenet i Valero, 2008), o bé sense cap estratificació (Aluja *et al.* 2008). Altres proporcionaran puntuacions estandarditzades derivades, en forma de puntuacions *T*, i de manera estratificada (Yueh-Hsien *et al.*, 2012). Les combinacions són evidentment múltiples, no són més que alguns exemples dels molts que poden ser reportats. Queda en mans de l'investigador proporcionar la informació necessària per a justificar una opció o una altra.

4. Equiparació de puntuacions

Equiparar les puntuacions de dues o més mesures consisteix a establir una correspondència entre les seves puntuacions. Aquesta estratègia ha de permetre que, recollida la puntuació en una de les mesures, sigui possible estimar la puntuació de les altres mesures.

Un dels contextos en què pot ser especialment interessant resoldre equiparacions entre tests és aquell en què és necessari mesurar les aptituds de persones de manera repetida i en què, per tant, es poden produir processos d'aprenentatge o memòria, o simplement cansament, que podrien conduir cap a una més que probable estimació esbiaixada de les capacitats de les persones. En aquestes condicions seria interessant poder disposar de dues o més formes alternatives de la mesura que en permetessin l'ús indistintament. Tanmateix, les condicions que s'hauran de satisfer per a una equiparació adequada són les següents:

- a) Que es mesuri la mateixa característica.
- b) Que les dues mesures siguin igualment fiables.
- c) Que la transformació sigui invertible, és a dir, la transformació ha de ser possible de A a B , però també de B a A .

Dos tipus de transformacions són possibles, l'equiparació horitzontal i la vertical. El primer cas s'esdevindria quan el tret d'interès hagués de mantenir la seva dificultat entre les diferents mesures alternatives. Un sol grup de subjectes és el que, generalment, proporciona les dues o més mesures. En el cas vertical, en canvi, s'assumeix que la dificultat entre formes alternatives és diferent. En aquesta darrera circumstància, són avaluats dos o més grups de persones que presenten capacitats diferenciades.

Els diversos dissenys que permeten crear equiparacions entre mesures són els següents:

- els de grup únic,
- els de grups equivalents,
- els tests d'ancoratge.

En el cas dels dissenys de **grup únic**, un mateix grup de persones extret aleatòriament de la població destinatària de les mesures equivalents és avaluat mitjançant les diverses mesures d'interès. Com a estratègia de control, per a evitar processos d'aprenentatge i memòria, i per tant un possible biaix com

a conseqüència d'haver administrat primer una forma alternativa i posteriorment l'altra, es recomana que les mostres responguin a les mesures de manera contrabalancejada. Així, es pot dividir la mostra aleatòriament en dos grups i administrar a cada subgrup les diverses formes alternatives en un ordre diferent, per exemple, en un grup la seqüència *A-B*, i en l'altre la *B-A*.

En els dissenys que impliquen **grups equivalents** són dos els grups que se seleccionen aleatòriament, però en aquest cas s'administra una mesura a la primera mostra, i a la segona mostra la segona mesura. Poder assumir amb garanties que les dues mostres de subjectes són comparables, per exemple, mitjançant una assignació aleatòria dels subjectes, és condició necessària per a assegurar la validesa del procediment.

L'avantatge d'emprar dos grups de persones és que, en comparació amb l'estratègia anterior, redueix l'efecte de cansament i no exposa els participants a processos d'aprenentatge o memòria que podrien ser arrossegats entre les diverses mesures, ja que els participants només s'exposen a una d'aquestes. La limitació més important d'aquesta aproximació és que, en general, exigeix incloure una quantitat de participants superior a la necessària en el cas del primer disseny.

Els **tests d'ancoratge** parteixen, també, de disposar de dos grups de participants en què cada un rep una de les mesures dirigides a ser equiparades. En aquest cas, però, a cada mostra se l'administra alguns ítems que formen part de la mesura que s'administra a l'altra mostra. Aquests ítems comuns esdevenen així ítems d'ancoratge. L'estratègia és especialment interessant quan no és possible garantir que els dos grups de persones siguin comparables, situació freqüent en un context aplicat en què no ha estat possible generar els dos grups de persones de manera aleatòria. Sota aquesta circumstància, els ítems d'ancoratge proporcionaran una estimació del grau en què les dues mostres esdevenen comparables.

Independentment de quina de les tres estratègies s'hagi emprat per a satisfer l'equivalència entre mesures, dos són els mètodes més comuns que es poden executar per a resoldre l'equiparació entre mesures: l'equipercentil i la transformació lineal.

L'**equiparació equipercentil**, que sol ser l'estratègia més habitual, consisteix a considerar equivalents les puntuacions directes que es vinculen a un mateix percentil. Així doncs, si les puntuacions directes 10 i 14 obtingudes amb la mesura *A* i *B*, respectivament, s'associen a un percentil 30, es considerarà que les dues puntuacions són equivalents. El mateix raonament seria aplicable a la resta de parells de puntuacions directes que convergissin en un mateix percentil.

Lectura recomanada

En el text de Muñiz (2003) trobareu més informació respecte a aquesta estratègia.

Les consideracions que s'han de tenir en compte a l'hora d'aplicar aquesta estratègia són les mateixes que ja s'han exposat en parlar dels percentils. Es tracta, doncs, d'una estratègia de fàcil comunicació i execució.

La **transformació lineal** consisteix a equiparar puntuacions estandarditzades. Concretament consisteix a equiparar les puntuacions directes que convergeixen en una mateixa puntuació típica. Expressat d'una altra manera, i d'acord amb la fórmula que permet calcular una puntuació típica z , es pot dir que:

$$\frac{X - \bar{X}}{S_x} = \frac{Y - \bar{Y}}{S_y}$$

D'on si s'aïlla Y :

$$Y = \frac{S_y}{S_x}(X - \bar{X}) + \bar{Y}$$

Exemple

Considerem l'exemple de dues mostres equivalents de persones extretes aleatòriament d'una determinada població (disseny de grups equivalents). La primera mostra obté en una mesura d'aptituds aritmètiques (X) una mitjana de 10 i una desviació típica de 4. La segona mostra, en la segona mesura també d'aptituds aritmètiques (Y), obté una mitjana de 8 i una desviació típica de 2. Segons aquesta informació, una puntuació de 9 en la mesura A , a quina puntuació directa correspondria en la mesura B ?

$$Y = \frac{2}{4}(9 - 10) + 8 = 7,5$$

Obtenir un 9 en la mesura A és equivalent a obtenir una puntuació de 7,5 en la mesura B .

En el cas del disseny d'un sol grup en el qual s'ha emprat una estratègia de contrabalanceig, cal tenir en consideració l'ordre en què es van administrar les dues mesures (ordre $A-B$ i ordre $B-A$). Sota aquesta circumstància es calcula la desviació típica i mitjana en cadascun dels grups de balanceig. Així doncs, començant per la desviació típica:

$$S_x = \sqrt{S_{x1}^2 + S_{x2}^2}$$

$$S_y = \sqrt{S_{y1}^2 + S_{y2}^2}$$

En què els estadístics inclosos en els radicals són les variàncies corresponents al primer i segon ordre en l'administració de les dues mesures.

En el cas de les mitjanes:

$$\bar{X} = \frac{\bar{X}_{x1} + \bar{X}_{x2}}{2}$$

$$\bar{Y} = \frac{\bar{X}_{y1} + \bar{X}_{y2}}{2}$$

Essent els estadístics dels numeradors les mitjanes corresponents al primer i segon ordre de les dues mesures. Calculades les mitjanes i desviacions típiques, només resta aplicar la fórmula ja coneguda que permet establir la correspondència entre les puntuacions directes equivalents de les dues mesures.

Exemple

Les mesures X i Y han estat administrades a dos grups de persones de manera contrabalancejada. La mesura X , quan es va administrar en primer lloc, va presentar una mitjana de 10 i una desviació típica de 2. Quan es va administrar en segon lloc, la mitjana va ser de 8 i la desviació típica d'1. En el cas de la mesura Y , administrada en primer lloc, va presentar una mitjana de 12 i una desviació típica de 3, mentre que quan es va administrar en un segon ordre la seva mitjana va ser de 14 i la desviació típica de 4. Quina puntuació correspondria a la mesura Y si en la mesura X s'obtingués una puntuació de 7?

Calculant les desviacions típiques de les mesures X i Y segons l'ordre en què es van administrar, respectivament:

$$2,24 = \sqrt{4 + 1}$$

$$5 = \sqrt{9 + 16}$$

En el cas de les mitjanes, respectivament:

$$9 = \frac{10 + 8}{2}$$

$$13 = \frac{12 + 14}{2}$$

Establint la correspondència entre les dues mesures, s'arribaria a la conclusió que la puntuació en la mesura Y hauria de ser 9:

$$\frac{5}{2,24}(7 - 9) + 13 = 8,54 \approx 9$$

L'aplicació d'aquesta forma d'equiparació entre mesures ha d'ajustar-se a les mateixes consideracions que es van recollir en parlar de les puntuacions estandarditzades. Implica, per tant, dipositar l'interès sobre la tendència de resposta dels individus en les dues mesures i assumir que les dues distribucions de valors objecte d'equiparació s'ajustin bàsicament a una distribució normal.

Resum

En la primera part d'aquest mòdul s'ha parlat dels percentils, les puntuacions estandarditzades i les normes cronològiques, puntuacions transformades comunament emprades, dirigides a atorgar interpretabilitat a les puntuacions obtingudes en una mesura. El percentils, calculats d'acord amb la posició d'una persona en un continu i expressats com a percentatges, es converteixen en una aproximació fàcil d'entendre i per tant de difondre, fins i tot a persones poc acostumades a l'ús d'aquest tipus d'estratègies matemàtiques. A més, el fet de ser essencialment insensibles als grups de valors no ajustats a una distribució normal permet que es puguin considerar com una transformació d'ampli espectre, una estratègia útil i aplicable per a la majoria de variables quantitatives que són d'interès en psicologia.

Les puntuacions estandarditzades presenten més limitacions en termes d'aplicabilitat. Requereixen que les observacions que són objecte d'anàlisi s'ajustin bàsicament a una distribució normal. I tal com s'ha recollit, són moltes les variables psicològiques que, avaluades en una mostra concreta de persones, acaben no satisfent aquesta forma de distribució. No obstant això, quan aquesta circumstància pot ser assumida, l'estandardització representa una aproximació diferenciada de la dels percentils que pot resultar especialment interessant. A diferència dels percentils, es focalitzen en la tendència de resposta, la mitjana de comportament que proporciona el grup normatiu. La desviació típica, el com de diferents són els individus també del grup normatiu, esdevé la manera de quantificar la discrepància entre la puntuació d'una persona concreta i aquest comportament mitjà.

Concloent el conjunt d'estratègies de transformació, s'ha exposat la puntuació en forma de normes cronològiques o edat. En aquest cas, i d'acord amb la tendència de resposta observada de nou en un grup normatiu, constituït per edats diferents, la puntuació d'un individu passa a ser transformada i interpretada d'acord amb aquesta variable cronològica. El seu raonament, com en el cas dels percentils, és també simple, fàcilment comunicable. No obstant això, i tal com s'ha fet notar, presenta inconvenients que no es poden menys tenir. Tot i aquestes limitacions, ha esdevingut una estratègia molt emprada en l'àmbit de la mesura de la intel·ligència, la qual cosa ha justificat l'espai dedicat en aquest mòdul.

El barem ha estat un dels conceptes fonamentals del mòdul. Gràcies a aquest, l'usuari d'un instrument pot fer inferències respecte a la puntuació d'un individu d'acord amb el perfil de comportament d'una mostra normativa, i emprar taules que estableixen aquesta correspondència, moltes vegades mitjançant l'ús de puntuacions transformades. No obstant això, més enllà de les consideracions que es deriven de l'ús d'una estratègia de transformació o altra,

l'element més rellevant que s'ha de tenir en consideració és la necessària i adequada justificació de quina ha de ser aquesta mostra normativa que estructuri aquest barem. Aquesta mostra determinarà fortament les inferències que se'n puguin derivar i podrà condicionar, quan s'estigui treballant en un context aplicat, el sentit de les decisions que s'hagin de prendre.

El mòdul conclou amb algunes de les estratègies emprades amb més freqüència dirigides a equiparar puntuacions entre mesures diferents que mesuren un mateix objecte psicològic. L'equiparació equipercentil o la transformació lineal esdevenen estratègies que poden ser especialment útils, per exemple, en els contextos de treball en què una avaluació repetida d'una mateixa variable en un mateix individu pot conduir a una estimació esbiaixada. Els avantatges i limitacions que l'usuari ha de tenir presents quan es vulgui emprar alguna d'aquestes dues estratègies són homòlegs dels descrits en el cas dels percentils o de les puntuacions estandarditzades.

Bibliografia

- Aiken, L. R. (1996). *Tests psicológicos y evaluación* (8a. ed.). Mèxic: Prentice Hall.
- Alegret, M., Espinosa, A., Vinyes-Junqué, G., Valero, S., Hernández, I., Tárraga, L., *et al* (2012). Normative data of a brief neuropsychological battery for Spanish individuals older than 49. *Journal of Clinical and Experimental Neuropsychology*, 34, 209-219.
- Aluja, A., Blanch, A., Solé, D., Dolcet, J. M., i Gallart, S. (2008). Validez convergente y estructural del NEO-PI-R. Baremos orientativos. *Boletín de Psicología*, 92, 7-25.
- American Educational Research Association, American Psychological Association, i National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Gomà-i-Freixanet, M. i Valero, S. (2008). Spanish normative data of the Zuckerman-Kuhlman Personality Questionnaire in a general population sample. *Psicothema*, 20, 324-330.
- Muñiz, J. (2003). *Teoría clásica de los tests*. Madrid: Pirámide.
- Ramos-Quiroga, J. A., Draigue, C., Valero, S., Bosch, R., Gómez-Barros, N., Nogueira, M., Palomar, G., Roncero, C., i Casas, M. (2009). Validación al español de la escala de cribado del trastorno por déficit de Atención/hiperactividad en adultos (ASRS v. 1.1): una nueva estrategia de puntuación. *Revista de Neurología*, 48, 449-452.
- Rodríguez, C., Jiménez, J. E., Díaz, A., García, E., Martín, R., i Fernández, S. (2012). Datos normativos para el Test de los Cinco Dígitos: desarrollo evolutivo de la flexibilidad en Educación Primaria. *European Journal of Education and Psychology*, 5, 27-38.
- Valero, S., Ramos-Quiroga, J. A., Gomà-i-Freixanet, M., Bosch, R., Gómez-Barros, N., Nogueira, M., Palomar, G. *et al*. (2012). Personality profile of adult ADHD: The alternative five factor model. *Personality Research*, 198, 130-134.
- Yueh-Hsien, L., Chwen-Yng, S., Wei-Yuan, G., i Yee-Pay, W. (2012). Psychometric validation and normative data of the second Chinese version of the Hooper Visual Organization test in Children. *Research in Developmental Disabilities*, 33, 1919-1927.
- Zuckerman, M., Kuhlman, D. M., Joireman, J., Teta, P., i Kraft, M. (1993). A comparison of the three structural models for personality: the big three, the big five, and the alternative five. *Journal of Personality and Social Psychology*, 65, 757-768.

Anàlisi dels ítems

PID_00216187

Albert Bonillo

Temps mínim de dedicació recomanat: 3 hores



**Albert Bonillo**

Professor lector de l'Àrea de Metodologia de les Ciències del Comportament de la Facultat de Psicologia de la Universitat Autònoma de Barcelona (UAB) i consultor dels Estudis de Psicologia i Ciències de l'Educació de la Universitat Oberta de Catalunya (UOC). Doctor en Psicologia per la UAB, màster en Disseny i estadística en ciències de la salut. En la seva tesi doctoral proposa un conjunt d'algorismes per depurar dades ja registrades. La seva recerca s'orienta, entre altres àmbits, a la depuració de dades, la detecció de plagi acadèmic, i als factors explicatius i conseqüències dels déficits en la funció executiva dels infants.

Índex

Introducció.....	5
1. Tipus de proves.....	7
1.1. Proves d'execució típica enfront de proves d'execució màxima	7
2. Directives en la construcció d'ítems.....	9
3. Teoria clàssica.....	12
3.1. Dificultat	12
3.2. Discriminació	15
3.3. Discriminació dels distractors	17
3.4. Valoració del biaix	20
4. Teoria de resposta a l'ítem.....	22
Resum.....	27
Bibliografia.....	29

Introducció

L'objectiu d'aquest mòdul és introduir l'estudiant en el tema de l'anàlisi d'ítems. Considerem un error que aquest tema no sigui present en tots els plans docents de l'assignatura de psicometria, ja que és prou important perquè valgui la pena estudiar-lo. És cert que, tradicionalment, l'estudi de la psicometria s'ha focalitzat més cap a l'estudi de les propietats dels instruments de mesura destinats a aspectes d'opinió i constructes psicològics, que cap als que mesuren coneixements o habilitats. No obstant això, el psicòleg de carrer treballa tant o més amb aquests últims que amb els primers.

Volem que l'estudiant sàpiga, des del principi, que aquest mòdul no el farà expert en cap dels aspectes que aborda. Li proporcionarà, ho desitgem i ho esperem, una bona introducció a cadascun dels temes, però és fàcil que per plaer –o necessitat professional, o totes dues coses– vulgui aprofundir en alguns dels aspectes estudiats. Recomanarem textos que sí els estudien en profunditat.

El mòdul s'inicia precisament distingint entre instruments en funció de l'objectiu. En segon lloc, i ja centrats en proves d'execució màxima, mostrarem quins són els aspectes que s'han de tenir en compte en la construcció dels seus ítems. En el tercer apartat veurem com s'analitzen les propietats psicomètriques de la prova i dels ítems a partir de la teoria clàssica de test (TCT). Veurem els conceptes de *difficultat* i *discriminació*, i aprendrem a valorar si un ítem és correcte o potser necessita una revisió. En el quart apartat, veurem una introducció a la teoria de resposta a l'ítem (TRI), que és una alternativa d'anàlisi a la TCT. Veurem la TRI de manera més succinta que la TCT. El model de TRI resol problemes teòrics de la TCT, però els càlculs d'aquesta última són més senzills i fàcilment aplicables que els de primera.

Deixarem per al final les conclusions que resumeixin tot el que hem presentat.

1. Tipus de proves

És tradicional que, quan des de l'àmbit de la psicologia parlem d'una prova –o d'un instrument de mesura–, pensem immediatament en una enquesta d'opinió, un test de personalitat o similars. Des del punt de vista del tipus de prova, aquestes que hem esmentat no són diferents del qüestionari de satisfacció sobre el servei que trobem a la sortida de molts hotels. Volen mesurar, en una persona, el valor determinat d'un constructe l'existència del qual es pressuposa.

Un cas diferent és una prova que vulgui ordenar els millors candidats a un lloc de treball. En aquest context, en què se suposa que hi ha un criteri –ser un bon treballador per al lloc ofert–, la mesura del constructe pot passar a un segon pla. L'objectiu de l'instrument és que cadascun dels ítems optimitzi la classificació correcta de les persones. Vegem, doncs, quines característiques tenen les proves en funció del que cerquen.

1.1. Proves d'execució típica enfront de proves d'execució màxima

Així doncs, i si classifiquem les proves per l'objectiu, en distingirem dos tipus bàsics. Anomenem *proves d'execució típica* –o *d'execució de trets*– les que mesuren aspectes no escalables o, dit d'una altra manera, les que plantegen preguntes que no tenen respostes ni correctes ni errònies, sinó que tracten d'aspectes d'opinió, de preferència o similars. Per contra, anomenem *proves d'execució màxima* les que avaluen constructes que sí són escalables, i que són aquelles en les quals té sentit parlar de respostes correctes o errònies. Un examen, un test d'intel·ligència o qualsevol instrument que mesuri l'aptitud es classificaria dins d'aquest epígraf.

Encara que tots els conceptes que hem vist fins ara en mòduls anteriors –*fiabilitat*, *validesa* i *transformació de les puntuacions obtingudes*– són aplicables a tots dos tipus d'instruments, les estratègies per a estudiar-los solen variar lleugerament i se solen estudiar aplicant-los a les proves d'execució típica. És cert que, per exemple, un test d'intel·ligència ha de ser fiable, però administrar-lo dues vegades en poques setmanes no pot tenir gaire sentit, ja que en l'interval els participants podrien haver obtingut la resposta correcta i contaminar així els resultats. No obstant això, sí té sentit repetir un test de personalitat amb pocs dies de diferència, i així comprovar si la mesura de l'instrument és tan estable com se suposa que és el constructe mesurat. En definitiva, les característiques que es volen estudiar depenen, és clar, de l'objectiu de l'instrument.

Sovint, el psicòleg professional no utilitza instruments estandarditzats, sinó que ha de crear ell mateix l'instrument. Si l'estudiant treballés en el departament d'RH d'una multinacional i aquesta li demanés una prova per a ocupar un lloc molt específic, què faria? Després de comprovar en el mercat que aquesta prova no existeix, hauria de crear-la. I hauria de fer-ho tenint en compte què es vol fer amb aquesta prova: seleccionar el millor treballador per a aquest lloc. Bé, i llavors què? Suposem que aquest lloc requereix certs coneixements. El psicòleg hauria de construir una prova que, a partir d'un nombre mínim d'ítems, pugui seleccionar el millor candidat.

Aprenuem, doncs, què s'ha de tenir en compte quan no hi ha més remei que crear una prova.

2. Directives en la construcció d'ítems

L'estudiant ja sap que l'objectiu principal d'aquest mòdul és mostrar com es mesura la qualitat d'un test de rendiment. Ara bé, no hem d'eludir l'explicació sobre què cal fer per a construir correctament una prova. Creiem que el treball d'un psicòleg no ha de ser únicament valorar si la prova està més o menys ben feta, sinó que també té molt a aportar en la seva construcció. Canviant totalment d'àmbit: no seria estrany que un arquitecte valorés edificis si no aprengué primer a construir-los?

Hi ha diversos treballs que exposen de manera molt exhaustiva quines són les directives que cal seguir per a construir correctament una prova d'execució màxima. Un dels primers i més coneguts és el de Haladyna, Downing i Rodríguez (2002). Es tracta, ni més ni menys, que de trenta-un criteris que s'han de seguir, classificats per apartats. Aquests criteris es refereixen al contingut de la pregunta (per exemple, cada ítem ha de mesurar un únic coneixement), al format, a l'estil (recomana ítems curts), a l'enunciat (evitar les negacions) i a les opcions de resposta (recomana evitar l'opció "Totes les anteriors són correctes/incorrectes").

Personalment, preferim els criteris de Moreno, Martínez i Muñiz (2004). Són menys (dotze), són molt més clars i més fàcils d'aplicar. Com podeu veure en la taula 1, ara els aspectes que s'han de valorar són tres: l'elecció del contingut, l'expressió i les opcions de resposta.

Taula 1. Noves directrius per a la construcció d'ítems d'elecció múltiple

A. Elecció del contingut que es vol avaluar

1. Ha de ser una mostra representativa del contingut recollit en una taula d'especificació, evitant ítems trivials.
2. La representativitat haurà de marcar el senzill o complex, concret o abstracte, memorístic o de raonament que hagi de ser l'ítem, com també la manera d'expressar-lo.

B. Expressió del contingut en l'ítem

3. El central s'ha d'expressar en l'enunciat. Cada opció és un complement que ha de concordar gramaticalment amb l'enunciat.
4. La sintaxi o estructura gramatical ha de ser correcta. S'han d'evitar ítems massa escarits o profusos, ambigus o confusos, cuidant a més les expressions negatives.
5. La semàntica ha d'estar ajustada al contingut i a les persones avaluades.

C. Construcció de les opcions

Font: pres de Moreno, Martínez i Muñiz (2004)

6. L'opció correcta ha de ser solament una, acompanyada per distractors plausibles.
7. L'opció correcta ha d'estar repartida entre les diferents ubicacions.
8. Les opcions han de ser preferiblement tres.
9. Les opcions s'han de presentar usualment en vertical.
10. El conjunt d'opcions de cada ítem ha d'aparèixer estructurat.
11. Les opcions han de ser autònomes entre elles, sense superposar-se ni referir-se les unes a les altres. Per això, s'han d'evitar les opcions "Totes les anteriors" i "Cap de les anteriors".
12. Cap opció no ha de destacar de la resta ni en contingut ni en aparença.

Font: pres de Moreno, Martínez i Muñiz (2004)

En el contingut, s'han de preguntar coses fonamentals. Sembla obvi, però, quants exàmens recordem en què se'ns preguntaven algunes qüestions que van aparèixer poc (o gens) a classe? Una prova hauria de contenir només (però tots) els conceptes fonamentals de l'assignatura que valora. La creença que en preguntar qüestions menors, en el fons, obliguem l'alumne a estudiar tota la matèria és absurda i afavoreix l'atzar. Respecte a l'atzar, recordem els alumnes que solament estudiaven mig programa –o menys– i confiaven en la sort el dia de l'examen.

Sobre l'expressió, les tres qüestions apuntades són òbvies, però de nou no sempre es compleixen.

Un exemple paradigmàtic de Moreno, Martínez i Muñiz (2004) mostra que és millor redactar aquest ítem:

En física es denomina *sublimació* un canvi de matèria:

1. Sòlida a gasosa
2. Líquida a sòlida
3. Gasosa a líquida

que no aquest:

En física, sublimació:

1. Representa un canvi de matèria sòlida a matèria gasosa.
2. Es refereix a un canvi de matèria líquida a matèria sòlida.
3. Consisteix en un canvi de matèria gasosa a matèria líquida.

Sobre les opcions de resposta, destacarem la recomanació que les opcions siguin independents entre elles, la qual cosa comporta automàticament no usar els cèlebres "Totes/Cap de les anteriors". És obvi que per a rebutjar una opció com "Totes les anteriors són correctes" només necessitem saber que una de les altres opcions no ho és. Així, d'un cop de ploma, podem eliminar dues opcions de les possibilitats i l'elecció es facilita molt. Si el test té tres opcions, ja coneixem la resposta, i si en té quatre, fins i tot ens podem arriscar a contestar a l'atzar entre les dues restants.

Per saber-ne més

Si l'estudiant vol aprofundir en aquest tema, recomanem acudir al text original de Moreno, Martínez i Muñiz (2004), en què, en un to molt didàctic i amb exemples molt accessibles, trobarà una explicació molt exhaustiva de cadascun dels criteris.

Exemple de prova d'execució màxima

A partir de les directrius que hem mostrat, i per il·lustrar amb un exemple concret i proper els conceptes que es presentaran en aquest mòdul, hem construït l'examen següent. Conté deu preguntes sobre aquest mateix mòdul i l'opció correcta està destacada en negreta.

1. La dificultat (ID) és un índex que indica la probabilitat...

- A. d'encertar-ho.
- B. de fallar-ho.
- C. de contestar-ho.

2. El valor de discriminació d'un ítem (ID) ha de ser...

- A. negatiu.
- B. diferent de 0.
- C. **positiu.**

3. Un distractor hauria de tenir discriminació...

- A. positiva.
- B. **negativa.**
- C. propera a 0.

4. Un test de personalitat és una prova...

- A. d'execució màxima.
- B. **d'execució típica.**
- C. de rendiment.

5. La fórmula per a calcular ID_c és...

- A. $\frac{A - \frac{E}{1-K}}{N}$.
- B. $\frac{A - \frac{K}{E-1}}{N}$.
- C. **$\frac{A - \frac{E}{K-1}}{N}$.**

6. El model de TRI calcula, a partir del coneixement,...

- A. la puntuació total esperada.
- B. **la probabilitat d'encertar un ítem.**
- C. la discriminació del test.

7. Els paràmetres a , b i c de la TRI indiquen, respectivament,...

- A. **discriminació, dificultat, pseudoendevinament.**
- B. dificultat, discriminació, pseudoendevinament.
- C. pseudoendevinament, discriminació, dificultat.

8. Si, per nivell de dificultat, solament poguéssim tenir ítems d'un tipus, aquests haurien de ser, generalment...

- A. fàcils.
- B. difícils.
- C. **mitjans.**

9. Un ítem que preguntí sobre un aspecte del temari difícil hauria de ser...

- A. fàcil.
- B. mitjà.
- C. **difícil.**

10. L'avaluació del biaix pretén...

- A. **fer més justes les proves.**
- B. avaluar la dificultat dels ítems.
- C. augmentar la fiabilitat de la prova.

Vegeu també

En l'últim apartat d'aquest mòdul es detalla un exemple de respostes (fictícies) d'un grup de vint alumnes a aquesta prova, al costat dels càlculs de la majoria d'índexs als quals fa rem referència en aquest text.

3. Teoria clàssica

Hi ha dues maneres principals d'apropar-se a l'anàlisi d'ítems. Distingirem, doncs, entre la teoria clàssica dels tests (TCT) i la teoria de resposta a l'ítem (TRI). La primera l'estudiarem en aquest apartat, i la segona en el següent. Quins supòsits té la TCT? Encara que s'estudia en profunditat en l'apartat de fiabilitat, es resumeixen en l'equació

$$X = V + E$$

Aquesta implica que la puntuació que una persona obté en contestar un instrument de mesura (X) conté l'anomenat *nivell vertader* d'aquesta persona (V) i una part d'error. L'objectiu de la TCT és, doncs, mesurar i minimitzar aquest error, la qual cosa implica analitzar la fiabilitat de la mesura. Com no podia ser d'una altra manera, i sempre en aquests supòsits, tots els indicadors de qualitat dels ítems depenen de la mostra de persones que els han contestat.

Vegem, ara, les propietats principals que s'han de mesurar d'un ítem.

3.1. Dificultat

L'índex de dificultat d'un ítem (ID) és la proporció de persones que el contesten correctament. És a dir,

$$ID = \frac{A}{N}$$

Lectura de la fórmula

A : nombre de persones que encerten l'ítem.

N : nombre total de persones que el contesten.

En tractar-se d'una proporció, ja que els encertants són un subconjunt dels que contesten, és obvi que els seus valors fluctuen entre 0 i 1, i sovint s'expressen com un percentatge. Paradoxalment, els valors propers a 1 indiquen una dificultat baixa –s'hauria d'anomenar, doncs, *índex de facilitat* i no *de dificultat*– i valors propers a 0 indiquen dificultat màxima. En la fila titulada ID de l'exemple de prova d'execució màxima del final del mòdul podem veure les dificultats dels ítems.

La fórmula anterior presenta un problema: no té en compte que una part dels encerts es donen per pur atzar. Com es tracta de preguntes amb alternatives tancades, és lògic pensar que una part dels encertants no coneixia la resposta i que si l'ha encertada és solament perquè ha escollit una de les alternatives però no perquè sàpiga la resposta. El problema que se'ns planteja és quants ho han fet? La solució és molt intuïtiva. Si 100 persones que no saben japonès contestessin un examen redactat en aquesta llengua amb preguntes de 4 alternatives, quantes preguntes esperaríem que encertessin? L'esperança matemà-

tica –i el sentit comú– ens diu que 25. I si hi hagués 5 alternatives de resposta? Òbviament, 20. Per tant, el nombre d'encerts total –i per tant, la probabilitat d'encertar una pregunta– depèn en certa manera del nombre d'alternatives de resposta.

Per tant, és recomanable utilitzar la fórmula següent:

$$IDc = \frac{A - \frac{E}{K-1}}{N}$$

Immediatament, veiem que la diferència entre totes dues fórmules és que en la segona es resta de A un nombre que s'obté de dividir els errors (E) entre el nombre d'alternatives errònies ($K - 1$).

Per aprehendre aquest concepte observem els resultats de l'ítem 4. L'han encertat 8 dels 20 participants. La seva dificultat sense corregir és, per tant, del 0,4. Ara bé, en corregir l'índex per l'encert a l'atzar $[12/(3-1)]/20 = 0,3$, obtenim un 0,3, això és, podem suposar que el 30% de les persones l'han encertat per casualitat. Val la pena observar que amb el primer càlcul obtindríem una dificultat del 40% (el podríem etiquetar com *de dificultat mitjana*), mentre que amb el segon obtindríem un $IDc = 0,1$ (dificultat alta):

$$IDc = \frac{8 - \frac{12}{3-1}}{20} = 0,1$$

Encara que conceptualment no té sentit, ja que continua essent una proporció, l' IDc pot ser inferior a 0: en aquest cas, s'assigna $IDc = 0$. Això és el que ocorre amb l'ítem 10, que té una dificultat corregida de -5%, cosa que no té sentit. Observem també que els ítems 1 i 3 són perfectament inútils, el primer per fàcil i el segon per difícil.

Un ítem que tots encerten o que tots fallen no serveix per a res més que per a perdre el temps contestant-lo. Si tothom qui respon l'encerta, és com si regaléssim a tots els alumnes una part de la puntuació. I si tots el fallen, és com si els penalitzéssim. Suposem una prova que té 10 ítems i una puntuació teòrica entre 0 i 10. En el primer cas que hem exposat, la puntuació real podria fluctuar entre 1 i 10, i en segon entre 0 i 9. És clar que això no parla bé de les propietats de la prova.

Una vegada sabem la dificultat d'un ítem, plantegem-nos el següent: com haurien de ser les dificultats de tots els ítems d'una prova? Com diu la directriu dos de Moreno, Martínez i Muñoz (2004), la dificultat d'un ítem s'ha de relacionar amb la del concepte que recull. Això és, si un contingut és fàcil, l'ítem ha de ser fàcil. Per tant, una prova que mesura continguts diversos hauria de tenir ítems de totes les dificultats, i aquestes s'haurien de correspondre amb la dificultat dels conceptes mesurats.

Lectura de la fórmula

A : nombre de persones que encerten l'ítem.

E : nombre de persones que fallen l'ítem.

K : nombre d'alternatives (o opcions) de resposta.

N : nombre total de persones que el contesten.

Una proposta, feta per nosaltres (Bonillo, 2012), és mostrar en un gràfic la dificultat (eix Y) dels ítems d'una prova (ordenats de menys a més dificultat en l'eix X) i observar si el pendent és de prop de 45°. Això és, la línia que uneix els punts d'aquestes dificultats hauria de creuar el gràfic en diagonal.

La figura següent mostra aquesta idea aplicada a ítems dels exàmens d'accés a la formació sanitària especialitzada (les cèlebres proves de metge, farmacèutic i infermer intern resident, MIR, FIR i IIR, respectivament) per a dos anys, el 2005 i el 2006. Si sumem totes dues convocatòries, s'hi van presentar més de 23.000 aspirants i els seus resultats són molt rellevants, ja que els aspirants que les superin seran els futurs metges, farmacèutics i infermers especialistes.

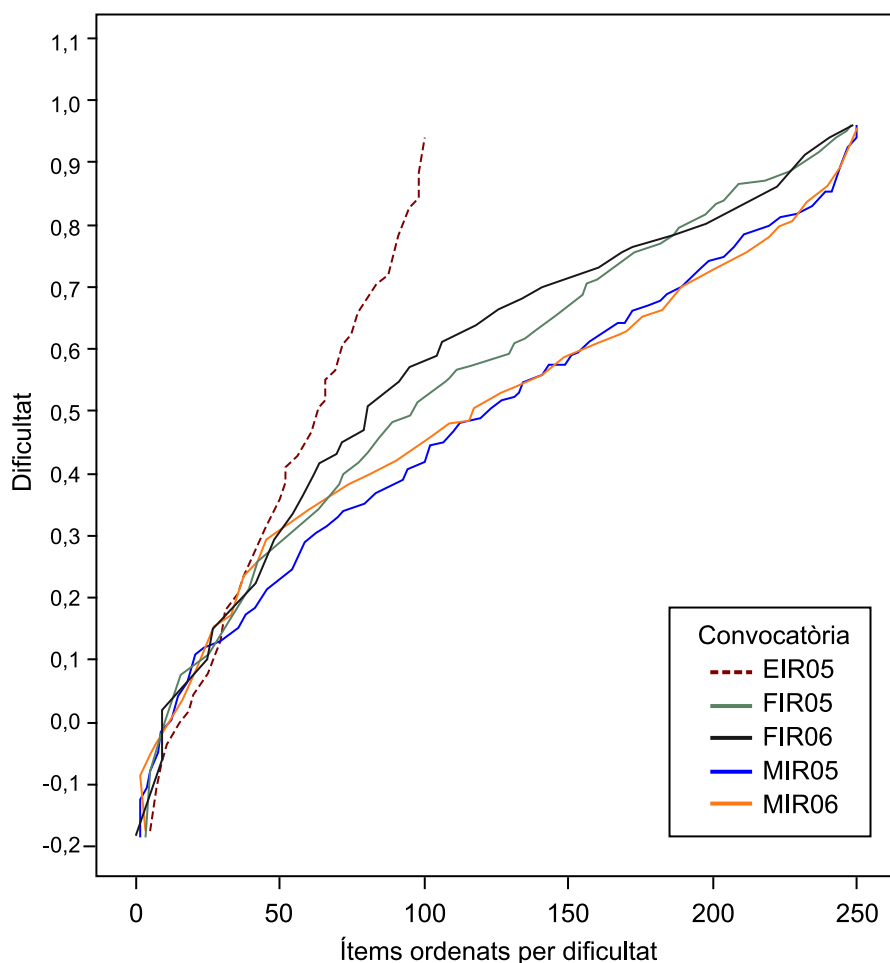


Figura 1. Ítems i dificultat de les proves d'FSE

Podem veure en el gràfic diverses qüestions que val la pena destacar. En primer lloc, veiem que hi ha dificultats negatives, i que aquestes s'expliquen perquè s'ha aplicat la correcció de l'atzar a ítems molt difícils. En segon lloc, cal tenir en compte que la prova d'infermeria consta de 100 ítems (enfront dels 250 de la resta de convocatòries), cosa que explica la diferència evident en les pendents. En tercer lloc, les corbes de les proves de farmàcia estan, en el gràfic, més altes que les de medicina, la qual cosa indica que són proves més fàcils. En quart lloc, les corbes són molt semblants intraprogrames, és a dir, la dificultat de les proves és molt semblant any rere any.

Per saber-ne més

Si l'estudiant vol conèixer amb més profunditat la proposta, que aquí solament apuntem, el remetem a l'article original: Bonillo (2012).

3.2. Discriminació

N'hi ha prou de saber si un ítem és fàcil o difícil per a decidir si és adequat o no? Intuïtivament, podríem pensar que sí, però estaríem equivocats. De fet, si haguéssim de destacar una propietat psicomètrica dels ítems sobre la resta, aquesta seria la discriminació. Si un ítem no discrimina, no és útil per al mesurament, i aquest és l'objectiu per al qual es va redactar.

Com el seu nom indica, entenem per *discriminació* la capacitat d'un ítem per a distingir entre les persones que tenen un bon rendiment en el test i els que tenen un mal rendiment.

Qui ha de contestar correctament una pregunta d'examen? No és tan important si són molts o pocs alumnes, com si els encertants són, en general, dels bons alumnes. A què ens referim quan diem "els bons"? Als alumnes que tenen una puntuació alta en la prova. És a dir, un ítem l'han d'encertar més els que han obtingut una puntuació alta en la prova que els que no l'han obtinguda. Òbviament, una pregunta no pot ser bona si solament l'encerten els pitjors alumnes: ha de passar el contrari.

L'índex de discriminació més popular és l'índex D, també anomenat *índex basat en les proporcions d'encerts*.

$$D = P_a - P_b$$

Les proporcions es calcularien com hem vist en la primera fórmula presentada, i, de nou, es poden expressar com a percentatges. Però, quins són els alumnes d'alt i baix rendiment? Hi ha diverses maneres de definir el punt de tall de la puntuació total en la prova per a fer aquesta classificació. D'una banda, és freqüent utilitzar la mediana de la puntuació total en la prova, cosa que crea dos grups igual de grans. Aquesta estratègia té com a avantatge que tots els participants participen en el càlcul, però té com a inconvenient clar que els grups són poc extrems. Intuïtivament veiem que dues persones amb un rendiment molt semblant poden estar en grups diferents solament per una petita diferència.

És preferible utilitzar grups més extrems per a poder estudiar correctament aquest índex. Kelley (1939) recomana utilitzar els percentils superior i inferior del 27%. Per què el 27% i no el 25%? Encara que l'article original demostra que el 27% és lleugerament millor que el 25%, en l'exemple amb respostes fictícies que mostrem s'utilitza el 25% com a criteri per a separar el grup de rendiment alt –s'interpretaria com el que obté puntuacions superiors al 75%

Lectura de la fórmula

P_a : proporció de persones del grup d'alt rendiment que encerta l'ítem.

P_b : proporció de persones del grup de baix rendiment que encerta l'ítem.

dels seus homòlegs– del baix –grup que reuneix el 25% de les puntuacions més baixes. Calcular el percentil 27 no sempre és senzill, i el 25 sí, i les variacions entre l'un i l'altre són molt més petites.

Quins són els límits del D? És obvi que, teòricament, poden fluctuar entre 1 i -1. El primer valor es donaria solament quan totes les persones del grup superior encertessin i totes les de l'inferior fallessin. El valor -1 solament es podria donar en el cas contrari, i llavors hauríem de dubtar sobre si la resposta considerada correcta ho és. Cap d'aquestes dues situacions se sol donar en la realitat.

Com hem d'interpretar, doncs, aquest índex? En primer lloc, només valors positius indiquen discriminació. És clar que un ítem ha de ser més encertat entre els millors. Però, quins valors indiquen una bona discriminació? Ebel i Frisbie (1991) van proposar la classificació següent, que s'ha de prendre com una orientació:

Taula 2. Punts de tall dels valors de discriminació (D) i la seva interpretació

D	Interpretació de la discriminació
0,40 o superior	Ítems amb discriminació molt bona.
0,30-0,39	Ítem amb una discriminació raonablement bona, però susceptible de ser millorat.
0,20-0,29	Es requereix millorar l'ítem.
<0,20	Discriminació pobre. L'ítem ha de ser eliminat o bé ha de ser revisat en profunditat.

Un motiu per a prendre la taula anterior amb precaució és que l'índex D depèn –i molt– de la dificultat. Si un ítem és molt difícil, tindrà pocs encertants (per definició) fins i tot en el grup d'alt rendiment. Si P_a és baixa, la D solament pot ser baixa. No sembla just comparar D d'ítems de dificultats molt diferents. Una alternativa proposada a l'índex D és calcular la diferència de proporcions relativa, en lloc de l'absoluta. És a dir,

$$Dr = (Pa - Pb) / P$$

Calculem la discriminació dels ítems 4 i 5, que tenen la mateixa dificultat.

$$\text{Ítem 4: } D = \left(\frac{4}{5} - \frac{0}{5} \right) \cdot 100 = 80\% \text{ i } Dr = \left(\frac{0,8 - 0}{0,8} \right) \cdot 100 = 100\%$$

$$\text{Ítem 5: } D = \left(\frac{2}{5} - \frac{0}{5} \right) \cdot 100 = 40\% \text{ i } Dr = \left(\frac{0,4 - 0}{0,4} \right) \cdot 100 = 100\%$$

Per a l'ítem 4, el primer valor ($D = 80\%$) s'hauria d'interpretar de la manera següent: els millors encerten l'ítem un 80% més que els pitjors. O interpretar-lo com una diferència de probabilitats: és un 80% més probable encertar l'ítem quan es té un rendiment alt (que si es té baix). El segon valor ($Dr = 100\%$) s'interpretaria com que els bons tenen un 100% més d'encerts (respecte als dolents). En aquest cas, tots dos valors parlen bé de la capacitat discriminativa de l'ítem. Per a l'ítem 5, les dades són semblants ($D = 40\%$ i $Dr = 100\%$). Així doncs, tots dos ítems serien més que acceptables.

Suposem ara un ítem molt difícil, que solament l'encerti 1 de 20 participants, i suposem que aquest pertany al grup d'alt rendiment. És clar que $D = 1/6 - 0/6 = 0,17 = 17\%$. Segons els criteris d'Ebel i Frisbie (1991) l'hauríem d'eliminar, però $Dr = 100\%$, i per tant té discriminació relativa màxima. Llavors, què hauríem de fer? No hi ha una resposta absoluta a aquesta pregunta.

Si, per les característiques de la prova, és acceptable tenir un ítem tan difícil, aquest s'hauria de mantenir, ja que discrimina tant com pot discriminar. Si, per contra, no és convenient que tan poques persones l'encertin, s'hauria de reformar, però la decisió –en aquest cas– depèn per complet de la dificultat i no de la discriminació.

En resum, cal tenir clar que la discriminació depèn de la dificultat i no s'ha d'interpretar *per se*. En termes estadístics, la discriminació depèn de la variància de la dificultat.

Hi ha altres índexs alternatius als que s'han presentat per a mesurar la discriminació. Un dels més utilitzats és la correlació ítem-test. Habitualment, s'utilitza l'índex de correlació biserial-puntual, ja que permet quantificar la relació entre una variable binària –encertar l'ítem o no– i una variable d'escala –la puntuació total de la persona en la prova, idealment sense tenir en compte l'ítem analitzat. La fórmula i la lògica d'aquesta prova es troba explicada àmpliament tant a Muñiz (2003)¹ com a Martínez Arias (1992)², però és molt senzill veure que una alta correlació –propera a 1– indica una gran discriminació de l'ítem, que valors propers a –1 indiquen el contrari (que els bons fallen l'ítem i els dolents l'encerten) i que valors propers a 0 indiquen que encertar aquest ítem no té res a veure amb el coneixement que mesura el conjunt de la prova.

⁽¹⁾Muñiz (2003, p. 220).

⁽²⁾Martínez Arias (1992, p. 556).

3.3. Discriminació dels distractors

Un aspecte clau per al bon funcionament d'un ítem és que els seus distractors³ realment ho siguin. Una alternativa que ningú –o gairebé ningú– no tria no confon les persones que responen i, per tant, no és útil. I al revés: una alternativa incorrecta que els millors trien molt potser no és tan incorrecta com va pensar qui la va redactar.

⁽³⁾*Distractors* és el nom que es dona a les alternatives de resposta incorrectes.

Com s'estudia el comportament dels distractors? L'índex habitual torna a ser l'índex D que ja hem vist, però en lloc de calcular-lo a partir dels qui encerten i els qui fallen, es fa a partir dels qui trien cadascuna de les alternatives de respostes.

Observem atentament els resultats d'administrar l'ítem 5. La discriminació de l'opció de resposta B és més alta que la de la resposta correcta, la C. Això s'ha d'interpretar com que els millors trien més un distractor –com és la B– que la resposta correcta –com és la C. Això implica que hem de revisar si, potser, la pauta conté un error. En el nostre cas no és així, i podem atribuir a l'atzar que l'opció C hagi estat tan escollida. Ara bé, potser hauríem de recalcar als alumnes què significa cada element de la fórmula (ja que és el que es pregunta en l'ítem 5) i enfortir així l'aprenentatge.

Quines propietats matemàtiques ha de tenir un distractor? Òbviament, tenir discriminació negativa, és a dir, ser més escollit entre els pitjors que entre els millors. A més, seria òptim que tots els distractors tinguessin una discriminació semblant, ja que indicaria que les seves capacitats d'atracció són semblants. Aconseguir això és especialment difícil, i aquesta dificultat creix exponencialment amb les alternatives de resposta. En resum: és molt més difícil redactar tres distractors que siguin efectius, que dos. Per això, la majoria d'estudis recomanen usar com a molt tres alternatives de resposta.

No hem de creure que les propietats que fins ara hem presentat són independents entre elles: res més allunyat de la realitat. Si un ítem té una opció de resposta inversemblant (per exemple, Maradona com a autor d'*El Quixot*), l'ítem serà més fàcil i necessàriament discriminarà pitjor.

Com ja hem fet quan hem estudiat la dificultat, ara ens plantegem si hi ha alguna manera d'estudiar el conjunt de les discriminacions dels ítems d'una prova.

De nou, la proposta és nostra (Bonillo, 2012) i consisteix a mostrar en un gràfic, anomenat *diagrama de caixa*, la discriminació de l'opció correcta i de cadascun dels distractors, ordenats de més grans a més petits.

La figura següent mostra això aplicat, de nou, als exàmens d'accés a la formació sanitària especialitzada.

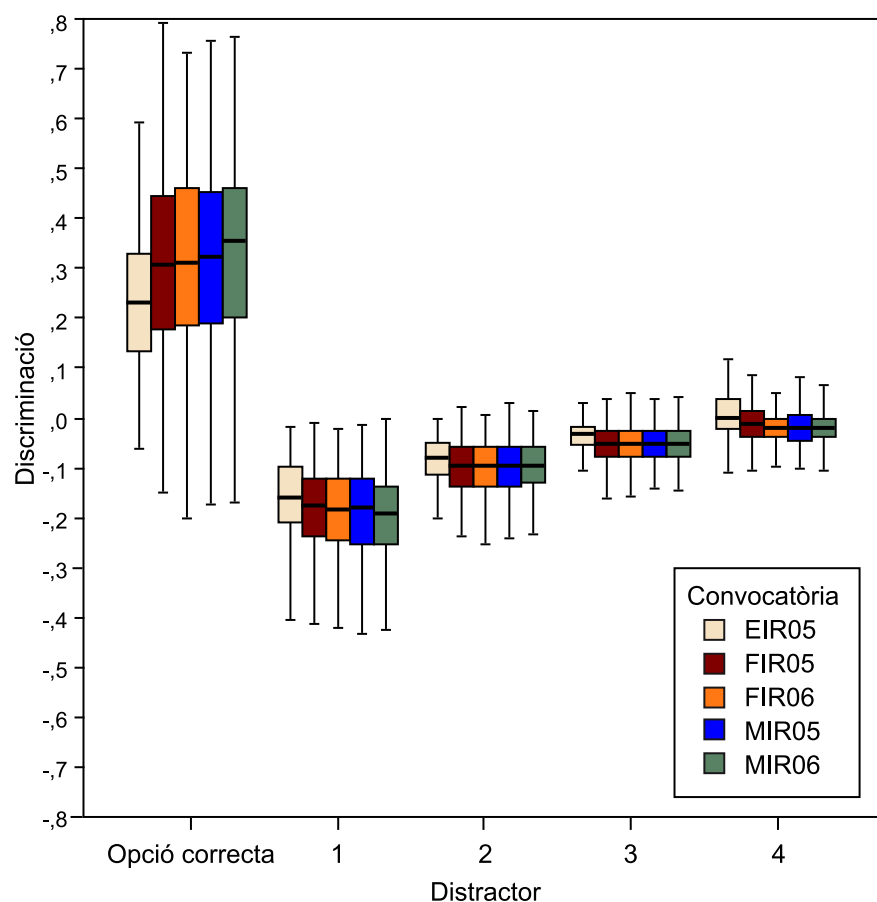


Figura 2. Discriminació dels distractors de les proves d'FSE

Cal tenir en compte que en aquesta figura apareixen les cinc convocatòries analitzades, això és $[250 \text{ ítems} \times 2 \text{ programes} \times 2 \text{ anys} + 100 \text{ ítems d'IIR}] \times 5 \text{ alternatives} = 5.500 \text{ valors}$. Així, i per a cada ítem, el distractor 1 és el més discriminatiu, i el 4 el menys. Com sol passar en aquests gràfics, les caixes mostren la mitjana –en traç gruixut– i els quartils –en els límits de les caixes. Les patilles (*whiskers*) mostren els valors mínims i màxims no allunyats ni extrems. Els allunyats es mostren amb punts i els extrems amb asteriscos.

S'observa que les discriminacions de les alternatives correctes són semblants entre especialitats i convocatòries. Destaquen de les altres les discriminacions relatives a la prova d'IIR, que són més baixes i menys disperses. En l'anàlisi dels distractors es veu que hi ha un escalat entre aquests, però que es redueix com més alternatives es tenen en compte; és a dir, la diferència entre la tercera i la quarta alternativa és molt més petita que entre la primera i la segona. També s'observa que les alternatives tres i quatre –recordem que són ordenades per la seva discriminació i que no s'han d'identificar amb alternatives de resposta D i E, per exemple– tenen discriminacions molt baixes o gairebé nul·les. Si considerem que el límit superior de les caixes de l'última alternativa és superior a 0, podem dir que més del 25% dels ítems té una alternativa de resposta amb discriminació positiva –és a dir, més escollida pel grup amb rendiment alt. A més, l'última alternativa presenta molts valors extrems i allunyats, això és,

Per saber-ne més

De nou, si l'estudiant vol conèixer a fons aquesta proposta ha de consultar l'article original (Bonillo, 2012), ja que en aquest mòdul no ens podem estendre molt més del que ja hem fet.

ítems en els quals, per la seva alta discriminació positiva, seria discutible si l'opció donada com a correcta veritablement ho és, o és l'única que ho és. Conclusió que n'hem de treure: amb tres opcions seria més que suficient.

3.4. Valoració del biaix

Un aspecte crucial quan es crea –i quan es valora– tant un ítem com un instrument de mesura és que aquests no siguin esbiaixats. De què parlem quan ens referim al biaix en un instrument de mesura? Una bàscula estarà esbiaixada si sempre infravalora el pes d'un objecte enfront d'un altre quan sabem que tots dos pesen exactament el mateix. En el context de les proves d'execució màxima, entenem que un ítem –o un test– està esbiaixat quan grups, per exemple homes i dones o rics i pobres, que tenen el mateix coneixement sobre la matèria mesurada no obtenen valors iguals, sinó que un dels grups és sistemàticament “perjudicat”.

Com us podeu imaginar, els instruments esbiaixats poden tenir greus implicacions socials. Si un examen, com la selectivitat, afavorís sistemàticament un alumne amb un nivell socioeconòmic alt enfront d'un que no té aquest nivell, sempre que sabéssim que tots dos saben exactament el mateix sobre el tema, la selectivitat no seria socialment justa.

Com es pot valorar el biaix? Actualment, s'utilitza el concepte de **funcionament diferencial dels ítems** i l'índex principal que se'n deriva és el DIF⁴. Diem que un ítem té o presenta DIF⁵ quan es donen diferències estadísticament significatives en la puntuació d'un ítem en dos grups diferents que haurien de tenir, per lògica, el mateix nivell. Per a avaluar el DIF es poden utilitzar diferents procediments matemàtics⁶, però en aquest text solament parlarem del mètode de Mantel-Haenszel (1959). Aquest és, en el marc de la TCT, el més utilitzat perquè és senzill de calcular i pels bons resultats que dona.

La manera de calcular i la idea que hi ha darrere el DIF és molt senzilla. Suposem que volem saber si la nostra prova no està afectada pel sexe dels que responen. És obvi que no ho hauria d'estar i que si no és així l'hauríem de millorar, ja que es podria qualificar de *sexista*.

Es tracta llavors de dividir els subjectes en grups en funció de les puntuacions totals (per exemple, en cinc grups). Després, calculem una taula per grup en què observem si la variable *sexe* s'associa a encertar més. Finalment, aquest resultat s'agrega a l'estadístic de Mantel-Haenszel i es compara amb el χ^2 de referència. Si el resultat és significatiu, vol dir que hi ha trams de puntuació total en què un sexe sembla que té avantatge sobre un altre, ja que té més encerts. Llavors, podem dir que la prova no és justa. Quan creem una prova

⁽⁴⁾La sigla DIF prové de les inicials del terme en anglès.

⁽⁵⁾En terminologia psicomètrica col·loquial, diem que hi ha DIF.

⁽⁶⁾A Muñiz (2003, pp. 239-253) se'n pot veure una excel·lent revisió.

Lectura recomanada

Recomanem encaridament la consulta de l'exemple que presenta Muñiz (pp. 247-249) i que aquí descriurem però no desenvoluparem matemàticament.

hauríem de comprovar que és independent de variables com el gènere, la raça i qualsevol altra que pugui comportar, implícitament, la discriminació de les persones.

4. Teoria de resposta a l'ítem

La teoria de resposta a l'ítem⁽⁷⁾ parteix d'una perspectiva totalment diferent. Critica la TCT en afirmar que estudia les propietats d'un test particular en una mostra –també particular– de persones. Des de la TCT, és cert que dos tests diferents però que mesuren el mateix constructe tindran propietats diferents. I també és cert que aquestes propietats dependran de si les persones utilitzades per a calibrar el test de rendiment tenen un rendiment alt o no. La TRI permet superar aquests problemes, però a costa de complicar enormement els càlculs i ser més difícil d'utilitzar en ser més “caixanegrista”.

(7) En anglès, *item response theory* (IRT).

La TRI es basa en el càlcul, per a cada ítem, d'una sèrie de paràmetres, que assumeixen un model matemàtic molt concret. L'objectiu últim és la mesura del tret latent⁽⁸⁾ a partir de tres paràmetres:

(8) *Tret latent* és el nom que rep el constructe que s'ha de mesurar, per exemple el coneixement d'una assignatura.

- la discriminació de l'ítem,
- la dificultat i
- l'encert a l'atzar.

(9) En anglès, *item response function* (IRF).

Aquests tres paràmetres es poden veure en la figura següent, que resulta clau per a entendre què és la TRI⁽⁹⁾: es coneix com la corba característica de l'ítem (CCI).

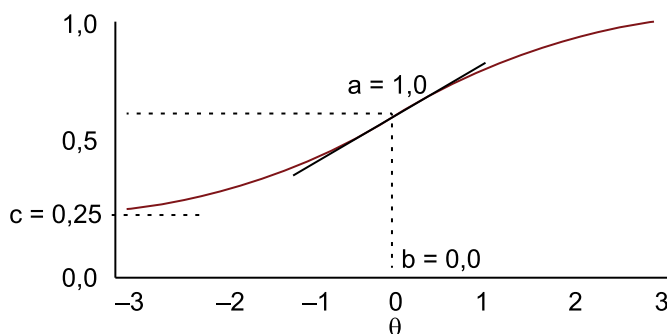


Figura 3. Exemple de corba característica d'un ítem (CCI)

La CCI mostra, en l'eix de les ordenades (l'eix Y), la probabilitat d'encertar l'ítem a partir de la magnitud del tret latent (eix d'abscisses o X). Aquesta probabilitat segueix una funció sigmoïdal (en forma de S, també anomenada *logística*) i el tret latent s'indica com a θ (theta). La funció logística té com a propietat que no pot ser més petit que 0 ni més gran que 1, i θ sol estar compresa entre -3 i +3.

A partir de la CCI podem mesurar els tres paràmetres clau de la TRI. El paràmetre indicat amb a) mesura la discriminació de l'ítem. Una corba molt plana expressaria que no és important tenir un alt coneixement del tret per a

augmentar la probabilitat d'encert. És a dir, com més gran sigui el pendent, més gran serà la discriminació. El del gràfic que es mostra és 1, valor positiu i acceptable.

El paràmetre indicat com a b mesura la dificultat a partir del punt de tall de l'eix X , que correspon a una probabilitat d'encert del 50%. S'interpretaria com el nivell de tret latent necessari per a tenir un 50% de probabilitat d'encertar l'ítem. Com més gran sigui la b , més difícil serà l'ítem, ja que caldrà més coneixement per a poder arribar a aquest 50% de probabilitat desitjat. El valor que es mostra en el gràfic és 0, que s'interpretaria com que és un ítem de dificultat (exactament) mitjana. El tercer paràmetre, indicat com a c , mesura el nivell d'atzar i també es coneix com a *índex de pseudoendevinament*. Gràficament, correspon al valor de X que talla l'eix Y . Recull, lògicament, la probabilitat d'encertar quan el coneixement de l'ítem és nul. El valor del gràfic indica que aquest és alt: del 25%.

Aquests càlculs es fan amb programari molt específic. Mostrar amb quines eines es fa i com, excedeix els objectius d'aquest mòdul.

Enllaç recomanat

La pàgina següent conté un conjunt de programes gratuïts que permeten aplicar la TRI: <http://www.psychology.gatech.edu/unfolding/FreeSoftware.html>. Ara bé, hem de recordar a l'estudiant que aplicar la TRI no és senzill i que requereix una formació més completa que la que necessita l'aplicació de la TCT.

Exemple d'avaluació de les propietats dels ítems a una mostra d'alumnes fictícia

En la taula següent es mostren les respostes dels 20 subjectes (per files) a les deu preguntes de la prova (columnes). En negreta es marquen les respostes correctes, i la seva pauta apareix en la primera fila.

Pauta	A	C	B	B	C	B	A	C	C	A
	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10
S1	A	C	C	B	C	B	A	C	C	B
S2	A	C	C	B	B	B	A	C	C	A
S3	A	C	C	B	B	B	A	C	C	A
S4	A	C	C	A	C	B	A	C	C	A
S5	A	C	A	B	B	B	A	C	C	A
S6	A	A	A	A	A	A	B	A	B	B
S7	A	C	A	B	C	A	B	B	B	B
S8	A	C	A	A	V	A	B	B	B	B
S9	A	A	A	B	C	A	B	B	B	C
S10	A	C	A	B	A	A	B	B	B	B

Pauta	A	C	B	B	C	B	A	C	C	A
	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10
S11	A	A	A	B	C	B	B	C	C	A
S12	A	A	A	A	A	B	B	C	C	A
S13	A	A	A	A	C	B	B	B	B	C
S14	A	A	A	A	C	B	B	B	B	B
S15	A	A	A	A	C	B	B	B	B	C
S16	A	A	A	A	B	C	B	B	C	B
S17	A	A	A	A	B	A	B	B	B	B
S18	A	A	A	A	A	C	B	A	A	C
S19	A	A	A	A	A	A	B	A	A	C
S20	A	A	A	A	A	A	B	A	A	C

La taula reproduïx l'anterior, però codificant la resposta com a 1 si s'ha encertat, i com a 0 si no s'ha encertat. La columna titulada "P. total" conté la puntuació total de cada persona, i és el resultat de sumar els uns de la taula. En la part inferior veiem dues files, titulades "ID" i "IDc". Com sabem, la primera correspon a l'índex de dificultat i la segona al mateix però en la versió corregida. Observeu la gran diferència entre tots dos i recordeu que el segon corregeix l'encert per atzar que el primer obvia.

	Encert										P. total
	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10	
S1	1	1	0	1	1	1	1	1	1	0	8
S2	1	1	0	1	0	1	1	1	1	1	8
S3	1	1	0	1	0	1	1	1	1	1	8
S4	1	1	0	0	1	1	1	1	1	1	8
S5	1	1	0	1	0	1	1	1	1	1	8
S6	1	0	0	0	0	0	0	0	0	0	1
S7	1	1	0	1	1	0	0	0	0	0	4
S8	1	1	0	0	0	0	0	0	0	0	2
S9	1	0	0	1	1	0	0	0	0	0	3
S10	1	1	0	1	0	0	0	0	0	0	3
S11	1	0	0	1	1	1	0	1	1	1	7
S12	1	0	0	0	0	1	0	1	1	1	5
S13	1	0	0	0	1	1	0	0	0	0	3
S14	1	0	0	0	1	1	0	0	0	0	3
S15	1	0	0	0	1	1	0	0	0	0	3

	Encert										P. total
S16	1	0	0	0	0	0	0	0	1	0	2
S17	1	0	0	0	0	0	0	0	0	0	1
S18	1	0	0	0	0	0	0	0	0	0	1
S19	1	0	0	0	0	0	0	0	0	0	1
S20	1	0	0	0	0	0	0	0	0	0	1
ID	100,0%	40,0%	0,0%	40,0%	40,0%	50,0%	25,0%	35,0%	40,0%	30,0%	
IDc	100,0%	10,0%	-50,0%	10,0%	10,0%	25,0%	-12,5%	2,5%	10,0%	-5,0%	

Els punts de tall dels quartils de la puntuació total són, respectivament, 1,75 i 8. Això és, quina puntuació mínima cal tenir per a estar per sobre del primer i del tercer quartil. Els valors compresos entre tots dos corresponen al grup anomenat *intermedi*, i que conté el 50% de les persones amb valors entorn de la mitjana. A partir d'aquests valors es calcula la columna que podem veure a la dreta de la taula, i que classifica les persones en tres grups.

	P. total	Grup rend.
S1	8	Superior
S2	8	Superior
S3	8	Superior
S4	8	Superior
S5	8	Superior
S6	1	Inferior
S7	4	Intermedi
S8	2	Intermedi
S9	3	Intermedi
S10	3	Intermedi
S11	7	Intermedi
S12	5	Intermedi
S13	3	Intermedi
S14	3	Intermedi
S15	3	Intermedi
S16	2	Intermedi
S17	1	Inferior
S18	1	Inferior
S19	1	Inferior
S20	1	Inferior

Les cel·les de la taula següent mostren els càlculs relatius a la discriminació de cadascun dels ítems. En les cel·les titulades “Encert del grup superior”, es mostra el percentatge de persones del grup de rendiment superior que s’ha escollit, respectivament, les alternatives A, B i C. Un càlcul idèntic es fa en les cel·les titulades “Encert del grup inferior”. Què fem després amb aquests valors?

Gràcies a aquest càlcul, podem obtenir de manera senzilla la discriminació de cadascuna de les opcions de resposta. En les cel·les titulades “Índex D alternatives” veiem la discriminació de l’alternativa A, i ídem per a les B i C amb les files inferiors. En negreta es mostra la discriminació de l’alternativa correcta. Observeu que la negreta coincideix amb la lletra que, en la primera taula d’aquest apartat, apareixia en la pauta de respostes correctes.

	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10
% encert grup superior										
A	100,0%	0,0%	20,0%	20,0%	0,0%	0,0%	100,0%	0,0%	0,0%	80,0%
B	0,0%	0,0%	0,0%	80,0%	60,0%	100,0%	0,0%	0,0%	0,0%	20,0%
C	0,0%	100,0%	80,0%	0,0%	40,0%	0,0%	0,0%	100,0%	100,0%	0,0%
% encert grup inferior										
A	100,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
B	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
C	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
Índex D alternatives										
A	0,0%	0,0%	20,0%	20,0%	0,0%	0,0%	100,0%	0,0%	0,0%	80,0%
B	0,0%	0,0%	0,0%	80,0%	60,0%	100,0%	0,0%	0,0%	0,0%	20,0%
C	0,0%	100,0%	80,0%	0,0%	40,0%	0,0%	0,0%	100,0%	100,0%	0,0%

Resum

Al llarg d'aquest text, hem exposat una multitud d'aspectes sobre els instruments de mesura. Sobre una prova hem vist què cal estudiar, com i quan s'ha de fer i hem insistit –especialment i conscientment– en el perquè cal fer-ho. La idea principal que ens agradaria haver transmès és que el que no podem fer és no fer res. És a dir, la pitjor de les situacions que podem imaginar –en aquest context, és clar– és aplicar una prova sense estudiar-ne cap de les propietats. No fer-ho convertirà l'error en norma i no en excepció.

Aquesta situació descrita és, lamentablement, la realitat. En l'àmbit universitari, que és el que ens és més proper, són pocs els professors que estudien els exàmens després d'administrar-los. Fins i tot els que ho fan, rares vegades publiquen els resultats, que permetrien als alumnes contrastar la justícia de la prova amb la qual se'ls va examinar. Per què passa això? Creiem que és més atribuïble a la falta de formació que no a la falta de transparència. El professor no és format en la manera com ha de valorar les seves proves, i tampoc no és format en la manera com ha de fer una classe. En el nostre sistema educatiu, i ens referim a totes les etapes, la cultura de treballar amb evidències no està implantada.

No podem no centrar-nos en l'àmbit educatiu, que és al qual pertanyem, però, no penseu que s'haurien de publicar les dades que permetin valorar la justícia d'unes oposicions, com són les proves d'accés de la formació sanitària especialitzada? De fet, en l'article que ja hem esmentat (Bonillo, 2012) aquesta va ser la nostra proposta principal. Així, els opositors podrien impugnar preguntes, proposar altres correccions i, en definitiva, tots podríem estar més tranquils en saber que les proves són justes i premien de debò els millors.

En aquest mòdul s'han presentat molts conceptes nous i s'ha fet de manera molt breu. Considerem que l'estudiant està preparat per a aplicar-los demà? En absolut. Ens agradaria pensar que, per a qui un dia necessiti crear o valorar una prova, aquest text és el primer d'altres. Aquests altres estan citats o se'n poden buscar alternatives.

Per a l'estudiant que no necessiti el que aquí s'exposa, que serà la gran majoria, confiem que els conceptes exposats s'entenguin. Quan això ocorre, i en el futur es necessita aplicar-los, recuperar-los de la memòria i de la biblioteca és senzill.

En qualsevol cas, i per a tots, esperem haver insuflat esperit crític i desig de treballar amb evidències i per les evidències. Al final, aquests dos elements són els que separen la psicologia d'altres coses que ens han de ser alienes.

Bibliografia

Baker, Frank (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation. Maryland: University of Maryland.

Bonillo, A. (2012). Pruebas de acceso a la formación sanitaria especializada para médicos y otros profesionales sanitarios en España: examinando el examen y los examinados. *Gaceta Sanitaria*, 26 (3), 231-235.

Downing S. M. (2005). The effects of violating standard item writing principles on test and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, 10, 133-143.

Ebel, R. L. i Frisbie, D. A. (1991). *Essentials of Educational Measurement*. NJ: Prentice-Hall.

Haladyna, T. M., Downing, S. M., i Rodríguez, M. C. (2002). A review of multiple-choice item-writing guidelines for Classroom Assessment. *Applied Measurement in Education*, 15 (3), 309-334.

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, 30 (1), 17-24. DOI: 10.1037/h0057123.

Mantel, N. i Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.

Martínez Arías, R. (1996). *Psicometría: Teoría de los Tests Psicológicos y educativos*. Madrid: Síntesis.

Moreno, R., Martínez, R., i Muñiz, J. (2004). Directrices para la construcción de ítems de elección múltiple. *Psicothema*, 16, 490-497.

Muñiz, J. (2003). *Teoría clásica de los tests*. Madrid: Pirámide.

Activitats pràctiques

PID_00216190

Antoni Cosculluela

Temps mínim de dedicació recomanat: 4 hores



**Antoni Cosculluela**

Professor titular d'universitat del Departament de Metodologia de les Ciències del Comportament de la Facultat de Psicologia de la Universitat de Barcelona (UB) i consultor dels Estudis de Psicologia i Ciències de l'Educació de la Universitat Oberta de Catalunya (UOC). Doctor en Psicologia per la UB. La seva tasca docent s'ha desenvolupat en l'àmbit de la metodologia i els dissenys de recerca, i de l'anàlisi de dades en psicologia i en informació i documentació. Autor i coautor de diferents llibres d'estadística aplicada. Actualment la seva recerca s'orienta, entre altres camps, als estudis bibliomètrics en diferents àrees de coneixement.

Índex

1. Introducció.....	5
2. Enunciats dels exercicis sobre fiabilitat.....	9
3. Enunciats dels exercicis sobre validesa.....	11
4. Enunciats dels exercicis sobre transformació i interpretació de les puntuacions.....	13
5. Enunciats dels exercicis sobre anàlisi dels ítems.....	14
6. Solucionari dels exercicis sobre fiabilitat.....	15
7. Solucionari dels exercicis sobre validesa.....	34
8. Solucionari dels exercicis sobre transformació i interpretació de les puntuacions.....	39
9. Solucionari dels exercicis sobre anàlisi dels ítems.....	46

1. Introducció

Per a fer una sèrie d'exercicis pràctics que posin en joc els continguts desenvolupats en aquest material, es proposen uns enuncisats que emmarquen uns casos generals, a partir dels quals es desenvoluparan activitats o exercicis concrets sobre continguts dels diferents mòduls vistos al llarg de l'assignatura.

Cas 1

S'han administrat dues formes paral·leles d'un mateix test de capacitat d'atenció compost per sis ítems cada forma, a una mostra de deu subjectes. El rang de valors de les puntuacions totals en el test va de 0 (capacitat d'atenció mínima) a 6 (capacitat d'atenció màxima). Es considera que un subjecte presenta una competència adequada en aquesta variable si obté una puntuació superior a dos punts.

Els resultats obtinguts han estat els següents:

Subjecte	x_1	x_2
1	3	2
2	5	5
3	0	2
4	6	5
5	1	0
6	3	3
7	4	2
8	5	6
9	2	3
10	1	3

x_1 i x_2 són les puntuacions dels subjectes en aquestes dues formes paral·leles.

Cas 2

Per a avaluar el nivell de depressió d'un subjecte, s'ha construït un test format pels sis ítems següents:

a) Em sento molt sovint trist i desanimat.

- b) Em considero una persona sense gaires problemes.
- c) No em sento decebut de mi mateix.
- d) Crec que he fracassat més que qualsevol persona normal.
- e) Normalment no tinc problemes per a dormir.
- f) Sovint obtinc força satisfaccions amb el que faig.

Les respostes (sí o no) dels subjectes es codifiquen amb un 0 o un 1 en funció de si presenten una tendència a la depressió o no. Així, la puntuació total en el test té un rang de valors que va de 0 (nivell de depressió mínim) a 6 (nivell de depressió màxim). Aquest test s'ha administrat a una mostra de vint-i-cinc subjectes, juntament amb l'inventari de depressió de Beck (BDI), que és un test àmpliament utilitzat i contrastat per a mesurar el grau de depressió i que consta de vint-i-un ítems. El rang de puntuacions en el BDI va de 0 (depressió mínima) a 21 (depressió màxima). Els resultats de les administracions d'aquests dos tests, amb les respostes dels subjectes als sis ítems del primer test, es presenten en la taula següent:

Test depressió								
Subjecte	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6	Total	BDI
1	1	1	0	0	0	1	3	10
2	1	1	1	0	1	1	5	13
3	0	1	1	0	0	0	2	8
4	0	1	1	0	0	0	2	12
5	0	0	1	1	1	1	4	16
6	0	0	0	1	1	1	3	15
7	0	0	1	0	1	1	3	12
8	0	0	1	0	0	0	1	10
9	1	1	1	0	1	1	5	19
10	0	0	1	0	1	1	3	13
11	0	1	1	0	0	1	3	16
12	1	1	0	1	1	1	5	13
13	1	0	1	1	1	1	5	20
14	0	0	0	0	0	0	0	3
15	1	0	0	0	1	1	3	12
16	1	0	1	0	0	0	2	9

Test depressió								
Subjecte	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6	Total	BDI
17	1	1	1	1	1	1	6	15
18	0	1	0	0	1	1	3	12
19	1	1	1	1	1	1	6	16
20	0	1	1	0	1	1	4	18
21	0	1	1	0	1	1	4	13
22	0	0	0	0	1	0	1	5
23	0	1	1	0	1	1	4	16
24	0	1	1	0	0	1	3	10
25	0	0	0	0	1	0	1	8

Cas 3

Amb l'objectiu d'avaluar el nivell d'intel·ligència general s'ha administrat un test de deu ítems a una mostra de vint-i-cinc subjectes. Els ítems són d'elecció múltiple amb quatre alternatives de resposta de les quals una és la correcta.

En la taula següent es presenten les respostes dels subjectes a aquest test, marcant amb negreta les respostes correctes de cada subjecte. Les alternatives de resposta correctes de cada ítem s'especifiquen en la darrera fila de la taula, si bé ja es podrien deduir de les respostes en negreta.

Subjecte	Ítems									
	1	2	3	4	5	6	7	8	9	10
1	A	C	D	A	B	A	B	D	A	D
2	B	C	A	A	B	C	B	D	C	B
3	C	C	C	B	A	B	B	A	B	C
4	A	C	D	A	B	C	B	D	D	A
5	D	A	A	B	B	D	B	B	B	D
6	D	B	A	C	C	D	A	A	D	D
7	B	C	C	A	B	A	B	D	C	D
8	C	C	A	A	B	B	B	A	C	D
9	C	C	D	A	B	C	B	D	A	C
10	B	C	A	C	D	A	C	C	D	D
11	B	D	C	D	A	D	B	B	B	A

* Alternativa de resposta correcta

Subjecte	Ítems									
12	A	C	D	A	B	C	B	B	C	C
13	C	C	C	B	B	B	B	D	B	A
14	D	C	D	A	B	B	B	D	C	D
15	B	C	B	C	A	D	B	C	D	B
16	A	C	D	A	B	C	B	D	B	D
17	D	C	A	A	B	A	B	D	A	D
18	D	A	A	D	B	D	B	A	A	A
19	B	C	D	A	B	B	B	D	A	D
20	B	C	A	A	B	D	B	B	C	C
21	C	C	C	D	C	B	B	D	A	D
22	C	D	C	B	D	A	A	C	A	D
23	C	C	D	A	B	A	B	D	A	B
24	A	C	D	A	B	C	B	D	B	D
25	B	C	D	A	B	B	B	D	C	A
*	A	C	D	A	B	C	B	D	A	D

* Alternativa de resposta correcta

2. Enunciats dels exercicis sobre fiabilitat

Exercici 1. A partir de l'enunciat del cas 1, resoleu les qüestions següents:

1. Obteniu el coeficient de fiabilitat del test.
2. Calculeu els indicadors següents d'acord entre les classificacions de les dues formes paral·leles:
 - a) Coeficient de Hambleton i Novick (p_{H-N}).
 - b) Coeficient kappa (k).
 - c) Coeficient de Livingston.

Exercici 2. Tenint en compte la informació i les dades del cas 2, contesteu els apartats següents en referència al test de depressió de sis ítems:

1. Obteniu la consistència interna (fiabilitat) del test utilitzant el mètode de les dues meitats a partir del següent:
 - a) La fórmula de Spearman Brown.
 - b) La fórmula de Rulon.
 - c) La fórmula de Gutman-Flanagan.
2. Obteniu el coeficient alfa (α) de Cronbach utilitzant les tres diferents fórmules per a calcular-lo, o sigui a partir del següent:
 - a) La variància dels diferents ítems.
 - b) La covariància entre els ítems.
 - c) $r1$.
3. Determineu la significació estadística del coeficient alfa calculat en l'apartat anterior (nivell de confiança del 95%), i el seu interval de confiança.
4. Si el mateix test s'administra a una mostra de trenta subjectes i s'obté un coeficient alfa de 0,70, hi ha diferències estadísticament significatives entre aquest coeficient alfa i l'obtingut en la mostra de vint-i-cinc subjectes del nostre exercici (nivell de confiança del 95%)?

5. En una segona administració del test a la mateixa mostra inicial de vint-i-cinc subjectes, s'ha obtingut un coeficient alfa de 0,64 i una correlació entre les puntuacions dels subjectes en aquestes dues aplicacions de 0,92. Hi ha diferències estadísticament significatives entre els valors dels dos coeficients alfa (nivell de confiança del 95%)?
6. Calculeu el KR_{20} i el KR_{21} de Kuder-Richardson. Compareu els dos valors i raoneu el perquè de la igualtat o la diferència.
7. Si afegíssim tres ítems més al test, suposant que mesuressin el mateix constructe, quina seria la seva nova fiabilitat?
8. Quants ítems hauríem d'afegir al test inicial per a arribar a una fiabilitat de 0,70?
9. Quina puntuació veritable podem estimar que tindrà el primer subjecte de la matriu de dades del nostre cas 1? Obteniu aquesta estimació amb un nivell de confiança del 95%, a partir del següent:
- a) La distribució normal dels errors.
 - b) El model de la regressió.

3. Enunciats dels exercicis sobre validesa

Exercici 1. A partir de les dades del cas 2, considerant les puntuacions del BDI com un criteri per a validar el nostre test de sis ítems, contesteu les qüestions següents:

- a) Calculeu el coeficient de validesa del test.
- b) Quin seria aquest coeficient de validesa si el test tingués una fiabilitat perfecta (tingueu en compte la fiabilitat del test obtinguda a partir del mètode de les dues meitats i la fórmula de Spearman Brown calculada en l'exercici 2, apartat 1.a)?
- c) Quan valdria el coeficient de validesa del test si hi afegíssim sis ítems més? Utilitzeu el mateix coeficient de fiabilitat de l'apartat anterior.
- d) Entre quines puntuacions en el criteri (BDI) podem pronosticar que obtindrà la puntuació un nou subjecte que té 4 punts en el nostre test? Construïu l'interval amb un nivell de confiança del 95%.
- e) Si suposem que el diagnòstic ja contrastat de trastorn depressiu lleu se situa en una puntuació superior a 15 en el BDI, i nosaltres volem comprovar la validesa de la decisió del nostre test de sis ítems, suposant que puntuacions superiors a 3 són les que detectarien aquest trastorn depressiu lleu, calculeu i interpreteu:

I. El percentatge d'acord entre els dos tests.

II. El coeficient kappa.

III. La sensibilitat del nostre test.

IV. L'especificitat del nostre test.

Exercici 2. Per tal de validar l'estructura interna del nostre test, hem fet una anàlisi de components principals (ACP) que ens ha proporcionat la matriu següent de saturacions factorials, en què s'han extret els dos components amb valor propi superior a 1:

Matriu de components

Component		
	1	2
Ítem_1	,587	,023

Component		
	1	2
Ítem_2	,267	,793
Ítem_3	,085	,703
Ítem_4	,696	–,303
Ítem_5	,705	–,360
Ítem_6	,821	,219

A partir d'aquestes dades que presenten les saturacions factorials de cada un dels sis ítems del test en els dos components extrets, calculeu les comunalitats dels ítems per a cada component i la conjunta per als dos components, el valor propi de cada component, i la variància explicada per a cada component i la conjunta dels dos components. Interpreteu els resultats respecte a l'estructura interna del test.

4. Enunciats dels exercicis sobre transformació i interpretació de les puntuacions

Exercici 1. Transformeu les puntuacions directes del test de depressió aplicat en el cas 2 a la mostra de vint-i-cinc subjectes, en les puntuacions següents:

- a) Percentils.
- b) Puntuacions estandarditzades.
- c) Puntuacions *T*.
- d) Enneatipus.
- e) Decatipus.

Exercici 2. En el cas 3 podem considerar que la puntuació total d'un subjecte en el test serà igual al nombre d'ítems que contesta correctament. D'altra banda, també podem considerar que, en ser un test d'intel·ligència, hi podem aplicar certes característiques de l'escala de Wechler per a adults o *Wechler adult intelligence scale* (WAIS).

Tenint en compte les consideracions anteriors, transformeu les puntuacions directes dels tres primers subjectes de la matriu de dades del cas 3 en percentils i en quocient intel·lectual (QI).

5. Enunciats dels exercicis sobre anàlisi dels ítems

Exercici 1. A partir de les dades del cas 3, calculeu i interpreteu els índexs de dificultat i de discriminació dels tres primers ítems del test. Per tal d'obtenir els índexs de discriminació considereu com a grup d'alt rendiment o alt nivell d'intel·ligència els subjectes que encerten més de set ítems, i de baix rendiment o baix nivell d'intel·ligència els que encerten menys de tres ítems. La grandària d'aquests dos grups no arriba al 27% o 25% del total dels vint-i-cinc subjectes, que és el recomanat, però són molt a prop d'aquests percentatges.

Exercici 2. Seguint amb les consideracions fetes en l'exercici anterior, feu l'anàlisi dels distractors del tres primers ítems.

6. Solucionari dels exercicis sobre fiabilitat

Exercici 1

1. En aquest cas, com que tenim les puntuacions d'una mostra de subjectes en les dues formes paral·leles d'un test, obtindrem el coeficient de fiabilitat aplicant el mètode de les formes paral·leles, i per tant hem de calcular el coeficient de correlació de Pearson entre les puntuacions dels subjectes en aquestes dues formes.

La fórmula del coeficient de correlació de Pearson és:

$$r_{xx'} = r_{x_1x_2} = \frac{n \sum x_1x_2 - \sum x_1 \sum x_2}{\sqrt{[n \sum x_1^2 - (\sum x_1)^2][n \sum x_2^2 - (\sum x_2)^2]}}$$

Així, per a obtenir-lo amb les dades de l'exercici, hem d'efectuar les operacions següents:

Subjecte	x_1	x_2	x_1x_2	x_1^2	x_2^2
1	3	2	6	9	4
2	5	5	25	25	25
3	0	2	0	0	4
4	6	5	30	36	25
5	1	0	0	1	0
6	3	3	9	9	9
7	4	2	8	16	4
8	5	6	30	25	36
9	2	3	6	4	9
10	1	3	3	1	9
Σ	30	31	117	126	125

$$r_{xx'} = r_{x_1x_2} = \frac{(10 \times 117) - (30 \times 31)}{\sqrt{[(10 \times 126) - (30)^2][(10 \times 125) - (31)^2]}} = 0,744$$

El coeficient de fiabilitat del test és de 0,744.

2. Per a calcular els coeficients d'acord en la classificació dels subjectes en les dues formes paral·leles d'aquest test, hem de generar la taula següent sabent que un subjecte és competent si té una puntuació superior a 2:

Subjecte	x_1	x_2	Classificació x_1	Classificació x_2
1	3	2	Competent	No competent
2	5	5	Competent	Competent
3	0	2	No competent	No competent
4	6	5	Competent	Competent
5	1	0	No competent	No competent
6	3	3	Competent	Competent
7	4	2	Competent	No competent
8	5	6	Competent	Competent
9	2	3	No competent	Competent
10	1	3	No competent	Competent

I a partir d'aquestes classificacions construïm la taula de contingència:

		x_2		
		Competents	No competents	
x_1	Competents	4	2	6
	No competents	2	2	4
		6	4	10

a) Coeficient de Hambleton i Novick (p_{H-N})

La fórmula del coeficient de Hambleton i Novich és:

$$p_{H-N} = p_c - p_a$$

En què, p_c és la proporció de classificacions consistents (6/10), i p_a la proporció de classificacions consistents que s'esperen per atzar i que s'obté amb la fórmula:

$$p_a = \sum \frac{n_{.j} \cdot n_{.i}}{N^2}$$

En què n_j és el nombre de subjectes classificats com a competents (o no competents) per la forma x_1 , n_i és el nombre de subjectes classificats com a competents (o no competents) per la forma x_2 , i N és el nombre total de subjectes.

$$p_a = \sum \frac{n_j \cdot n_i}{N^2} = \frac{6 \cdot 6}{10^2} + \frac{4 \cdot 4}{10^2} = 0,52$$

$$p_{H-N} = p_c - p_a = 0,60 - 0,52 = 0,08$$

Podem concloure que l'acord en les classificacions a partir de les dues formes paral·leles del test només aconsegueix un 8% més que l'acord previst per pur atzar. Cal afegir que aquest resultat tan pobre és, en bona part, degut a l'escassa grandària de la mostra.

b) Coeficient kappa (k)

Per obtenir el coeficient k , apliquem la fórmula següent:

$$k = \frac{p_c - p_a}{1 - p_a}$$

En què p_c i p_a són respectivament la proporció de subjectes classificats de manera consistent i la que s'esperaria per atzar tal com s'ha definit abans.

Per al nostre cas:

$$k = \frac{p_c - p_a}{1 - p_a} = \frac{0,60 - 0,52}{1 - 0,52} = 0,17$$

També en aquest cas, igual que en l'apartat anterior, el coeficient k de 0,17 indica una consistència molt pobre en la classificació, en bona part deguda als pocs subjectes de la mostra.

c) Coeficient de Livingston

El coeficient de Livingston l'obtenim a partir de l'expressió següent:

$$K_{xx'}^2 = \frac{r_{xx} S_x S_{x'} + (\bar{x}_x - C)(\bar{x}_{x'} - C)}{\sqrt{[S_x^2 + (\bar{x}_x - C)^2][S_{x'}^2 + (\bar{x}_{x'} - C)^2]}}$$

En què:

r_{xx} és el coeficient de fiabilitat a partir del procediment de formes paral·leles o test-retests.

S_x i $S_{x'}$ corresponen respectivament a la desviació típica del test en la primera i segona administració o en cada una de les formes paral·leles del test.

$\bar{x}_x$ i $\bar{x}_{x'}$ corresponen respectivament a la mitjana del test en la primera i segona administració o en cada una de les formes paral·leles del test.

C és el punt de tall.

S_x^2 i $S_{x'}^2$ corresponen respectivament a la variància del test en la primera i segona administració o en cada una de les formes paral·leles del test.

Per al nostre cas, aquests valors són:

$$r_{xx} = 0,744$$

$$S_x = 1,9$$

$$S_{x'} = 1,7$$

$$\bar{x}_x = 3$$

$$\bar{x}_{x'} = 3,1$$

$$C = 2$$

$$S_x^2 = 3,6$$

$$S_{x'}^2 = 2,89$$

I el coeficient de Livingston serà:

$$K_{xx'}^2 = \frac{r_{xx} S_x S_{x'} + (\bar{x}_x - C)(\bar{x}_{x'} - C)}{\sqrt{[S_x^2 + (\bar{x}_x - C)^2]} \sqrt{[S_{x'}^2 + (\bar{x}_{x'} - C)^2]}} = \frac{0,744 \cdot 1,9 \cdot 1,7 + (3 - 2)(3,1 - 2)}{\sqrt{[3,6 + (3 - 2)^2]} \sqrt{[2,89 + (3,1 - 2)^2]}} = 0,81$$

El coeficient de Livingston ens proporciona un valor de 0,81, que es pot considerar com una consistència adequada en les classificacions de les dues formes paral·leles.

Exercici 2

1.

a) Fórmula de Spearman-Brown

Obtenim la consistència interna o fiabilitat del test utilitzant el mètode de les dues meitats, a partir de la fórmula de Spearman-Brown, seguint aquests passos:

En primer lloc calculem, per a cada subjecte, el sumatori dels ítems parells i el dels ítems imparells:

Subjecte	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6	Parells	Imparells
1	1	1	0	0	0	1	2	1
2	1	1	1	0	1	1	2	3
3	0	1	1	0	0	0	1	1
4	0	1	1	0	0	0	1	1
5	0	0	1	1	1	1	2	2
6	0	0	0	1	1	1	2	1
7	0	0	1	0	1	1	1	2
8	0	0	1	0	0	0	0	1
9	1	1	1	0	1	1	2	3
10	0	0	1	0	1	1	1	2
11	0	1	1	0	0	1	2	1
12	1	1	0	1	1	1	3	2
13	1	0	1	1	1	1	2	3
14	0	0	0	0	0	0	0	0
15	1	0	0	0	1	1	1	2
16	1	0	1	0	0	0	0	2
17	1	1	1	1	1	1	3	3
18	0	1	0	0	1	1	2	1
19	1	1	1	1	1	1	3	3
20	0	1	1	0	1	1	2	2
21	0	1	1	0	1	1	2	2
22	0	0	0	0	1	0	0	1
23	0	1	1	0	1	1	2	2
24	0	1	1	0	0	1	2	1
25	0	0	0	0	1	0	0	1

En segon lloc apliquem la fórmula del coeficient de correlació de Pearson entre el sumatori dels ítems parells i el dels imparells, i n'obtenim un coeficient de correlació igual a 0,547 ($r_{pi} = 0,547$).

En tercer lloc apliquem la fórmula de Spearman-Brown per obtenir el coeficient de fiabilitat del test:

$$r_{xx'} = \frac{2r_{pi}}{1+r_{pi}} = \frac{2 \times 0,547}{1+0,547} = 0,707$$

El coeficient de fiabilitat o consistència interna del test aplicant la fórmula de Spearman-Brown és de 0,707.

b) Fórmula de Rulon

Com sabem, la fórmula de Rulon és:

$$r_{xx'} = 1 - \frac{S_d^2}{S_x^2}$$

En què:

S_d^2 : variància de les diferències entre les puntuacions dels subjectes en les dues meitats del test.

S_x^2 : variància de les puntuacions totals dels subjectes en el test.

Per tant, necessitem les puntuacions totals dels subjectes en el test, les seves puntuacions en els ítems parells i en el senars, i calcularem les diferències entre aquestes puntuacions (D).

Subjecte	Total	Parells	Imparells	D
1	3	2	1	1
2	5	2	3	-1
3	2	1	1	0
4	2	1	1	0
5	4	2	2	0
6	3	2	1	1
7	3	1	2	-1
8	1	0	1	-1
9	5	2	3	-1
10	3	1	2	-1

Subjecte	Total	Parells	Imparells	D
11	3	2	1	1
12	5	3	2	1
13	5	2	3	-1
14	0	0	0	0
15	3	1	2	-1
16	2	0	2	-2
17	6	3	3	0
18	3	2	1	1
19	6	3	3	0
20	4	2	2	0
21	4	2	2	0
22	1	0	1	-1
23	4	2	2	0
24	3	2	1	1
25	1	0	1	-1

Amb aquestes dades, calcularem les variàncies d'aquestes diferències i de la puntuació total.

Per exemple, en el cas de les puntuacions de diferència, aquesta variància serà igual a:

$$S_d^2 = \frac{\sum (D - \bar{D})^2}{n}$$

Sabent que la mitjana de les puntuacions de diferència és igual a -0,20, obtindrem la variància amb les dades següents:

D	$D - \bar{D}$	$(D - \bar{D})^2$
1	1,2	1,44
-1	-0,8	0,64
0	0,2	0,04
0	0,2	0,04
0	0,2	0,04
1	1,2	1,44

D	$D - \bar{D}$	$(D - \bar{D})^2$
-1	-0,8	0,64
-1	-0,8	0,64
-1	-0,8	0,64
-1	-0,8	0,64
1	1,2	1,44
1	1,2	1,44
-1	-0,8	0,64
0	0,2	0,04
-1	-0,8	0,64
-2	-1,8	3,24
0	0,2	0,04
1	1,2	1,44
0	0,2	0,04
0	0,2	0,04
0	0,2	0,04
-1	-0,8	0,64
0	0,2	0,04
1	1,2	1,44
-1	-0,8	0,64
	Σ	18

I, per tant, la variància de les diferències serà:

$$S_d^2 = \frac{\sum (D - \bar{D})^2}{n} = \frac{18}{25} = 0,72$$

Aplicant la mateixa fórmula a les puntuacions totals, obtindrem una variància de 2,4224.

Amb aquestes dades ja podem calcular el coeficient de fiabilitat a partir de la fórmula de Rulon:

$$r_{xx'} = 1 - \frac{S_d^2}{S_x^2} = 1 - \frac{0,72}{2,4224} = 0,703$$

El coeficient de fiabilitat o consistència interna del test aplicant la fórmula de Rulon és de 0,703.

c) Fórmula de Gutman-Flanagan

La fórmula de Gutman-Flanagan calcula el coeficient de fiabilitat d'un test a partir de les variàncies dels ítems parells i imparells, i de la puntuació total. La seva expressió és:

$$r_{xx'} = 2 \left(1 - \frac{S_p^2 + S_i^2}{S_x^2} \right)$$

En què:

S_p^2 : variància de les puntuacions dels subjectes en els ítems parells del test.

S_i^2 : variància de les puntuacions dels subjectes en els ítems imparells del test.

S_x^2 : variància de les puntuacions totals dels subjectes en el test.

Per al nostre exercici, calcularem les variàncies anteriors, i obtindrem els resultats següents:

$$S_p^2 = 0,8890$$

$$S_i^2 = 0,6816$$

$$S_x^2 = 2,4224$$

I la fórmula de Gutman-Flanagan ens proporciona el resultat següent:

$$r_{xx'} = 2 \left(1 - \frac{S_p^2 + S_i^2}{S_x^2} \right) = 2 \left(1 - \frac{0,8896 + 0,6816}{2,4224} \right) = 0,703$$

El coeficient de fiabilitat o consistència interna del test aplicant la fórmula de Gutman-Flanagan és de 0,703.

2.

a) Coeficient alfa de Cronbach a partir de la variància dels diferents ítems

Per obtenir el coeficient alfa a partir de les variàncies dels diferents ítems, aplicarem la fórmula següent:

$$\alpha = \frac{n}{n-1} \left[1 - \frac{\sum_{j=1}^n S_j^2}{S_x^2} \right]$$

En què:

n = nombre d'ítems del test.

$\sum_{j=1}^n S_j^2$: sumatori de les variàncies dels n ítems.

S_x^2 = variància de les puntuacions totals en el test.

Per tant, hem de calcular les variàncies de cada ítem i de les puntuacions totals en el test que ja hem calculat anteriorment.

	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6
Variàncies (S_j^2)	0,2304	0,2464	0,2176	0,1824	0,2176	0,2016

I aplicant-ho a la fórmula del coeficient alfa:

$$\alpha = \frac{n}{n-1} \left[1 - \frac{\sum_{j=1}^n S_j^2}{S_x^2} \right] = \frac{6}{5} \left(1 - \frac{0,2304 + 0,2464 + 0,2176 + 0,1824 + 0,2176 + 0,2016}{2,4224} \right) = 0,558$$

El coeficient alfa del test és de 0,558.

b) Coeficient alfa de Cronbach a partir de la covariància entre els ítems

La fórmula per a obtenir el coeficient alfa a partir de les covariàncies entre els ítems és:

$$\alpha = \frac{n}{n-1} \left[\frac{\sum_{j \neq k}^n \text{cov}(j, k)}{S_x^2} \right]$$

En què:

n = nombre d'ítems del test.

$\sum \sum \text{cov}(j, k) =$: sumatori de les covariàncies dels n ítems.

$S_x^2 =$: variància de les puntuacions en el test.

Com ja sabem, la covariància entre dues variables és igual a:

$$\text{cov}_{xy} = \frac{(x - \bar{x})(y - \bar{y})}{n}$$

Aquesta fórmula, aplicada a la covariància entre l'ítem 1 (x) i l'ítem 2 (y), ens donarà:

Ítem 1: x	Ítem 2: y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
1	1	0,64	0,44	0,2816
1	1	0,64	0,44	0,2816
0	1	-0,36	0,44	-0,1584
0	1	-0,36	0,44	-0,1584
0	0	-0,36	-0,56	0,2016
0	0	-0,36	-0,56	0,2016
0	0	-0,36	-0,56	0,2016
0	0	-0,36	-0,56	0,2016
1	1	0,64	0,44	0,2816
0	0	-0,36	-0,56	0,2016
0	1	-0,36	0,44	-0,1584
1	1	0,64	0,44	0,2816
1	0	0,64	-0,56	-0,3584
0	0	-0,36	-0,56	0,2016
1	0	0,64	-0,56	-0,3584
1	0	0,64	-0,56	-0,3584
1	1	0,64	0,44	0,2816
0	1	-0,36	0,44	-0,1584
1	1	0,64	0,44	0,2816
0	1	-0,36	0,44	-0,1584
0	1	-0,36	0,44	-0,1584
0	0	-0,36	-0,56	0,2016
0	1	-0,36	0,44	-0,1584

Ítem 1: x	Ítem 2: y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
0	1	-0,36	0,44	-0,1584
0	0	-0,36	-0,56	0,2016
$\bar{x} = 0,36$	$\bar{y} = 0,56$		Σ	0,96

$$\text{cov}_{xy} = \frac{(x - \bar{x})(y - \bar{y})}{n} = \frac{0,96}{25} = 0,0384$$

Aplicant la mateixa fórmula a la resta de covariàncies, tenim la matriu de covariàncies següent:

	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6
Ítem 1		0,0384	-0,0048	0,0736	0,0352	0,0608
Ítem 2	0,0384		0,0592	-0,0144	-0,0208	0,0768
Ítem 3	-0,0048	0,0592		-0,0032	-0,0224	0,0304
Ítem 4	0,0736	-0,0144	-0,0032		0,0768	0,0672
Ítem 5	0,0352	-0,0208	-0,0224	0,0768		0,1104
Ítem 6	0,0608	0,0768	0,0304	0,0672	0,1104	

I el sumatori d'aquestes covariàncies serà igual a:

$$\sum \sum \text{cov}(j, k) = 0,0384 - 0,0048 + 0,0736 + \dots + 0,0768 + 0,0672 + 0,1104 = 1,264$$

mentre que la variància de les puntuacions totals (S_x^2) ja sabem que és igual a 2,4224.

Per tant:

$$\alpha = \frac{n}{n-1} \left[\frac{\sum \sum_{j \neq k}^n \text{cov}(j, k)}{S_x^2} \right] = \frac{6}{5} \left(\frac{1,264}{2,4224} \right) = 0,558$$

El coeficient alfa del test és de 0,558, com ja hem vist en l'apartat anterior.

c) Coeficient alfa de Cronbach a partir de r_1

La fórmula del coeficient alfa també es pot expressar en funció del quocient entre la mitjana de les covariàncies i la mitjana de les variàncies dels diferents ítems del test. Aquest quocient, que designen com a r_1 , constitueix una estimació de la fiabilitat de cada ítem. En aquest sentit, la fórmula del coeficient

alfa a partir de r_1 és una aplicació de la correcció de Spearman-Brown, que hem comentat per al cas de les dues meitats, a partir de l'estimació de la fiabilitat de cada ítem, tenint en compte que si tenim n ítems és com si haguéssim allargat n cops l'ítem inicial.

La fórmula del coeficient alfa a partir de r_1 és la següent:

$$\alpha = \frac{n(r_1)}{1 + (n-1)r_1}$$

En què r_1 és el quocient entre la mitjana de les covariàncies i la mitjana de les variàncies dels diferents ítems del test, i n és el nombre d'ítems del test.

Per al nostre exercici, la mitjana de les covariàncies és $1,1264/30$, mentre que la mitjana de les variàncies és $1,296/6$.

Per tant,

$$r_1 = \frac{1,1264/30}{1,296/6} = 0,1738$$

I el valor d'alfa:

$$\alpha = \frac{n(r_1)}{1 + (n-1)r_1} = \frac{6 \times 0,1738}{1 + (5 \times 0,1738)} = 0,558$$

Com podem observar, i com no podia ser d'una altra manera, és exactament igual al valor obtingut en els dos apartats anteriors.

3. Significació estadística del coeficient alfa calculat en l'apartat anterior i determinació del seu interval de confiança

Per a comprovar la significació estadística d'un coeficient alfa, hem de seguir una sèrie de passos:

- Plantejament de la hipòtesi nul·la i l'alternativa:
 - Hipòtesi nul·la: $\alpha = 0$
 - Hipòtesi alternativa: $\alpha \neq 0$
- Càlcul de l'estadístic de contrast:

$$F = \frac{1 - \alpha}{1 - \hat{\alpha}} = \frac{1 - 0}{1 - 0,558} = 2,262$$

que es distribueix segons una F de Snedecor amb $(N - 1)$ i $(N - 1)(n - 1)$ graus de llibertat, essent:

- N : nombre de subjectes.
- n : nombre d'ítems.
- α : valor d'alfa en la població.
- $\hat{\alpha}$: valor d'alfa calculat en la mostra.
- Valors crítics de la distribució F de Snedecor amb 24 ($N - 1$) i 120 ($(N - 1)(n - 1)$) graus de llibertat, per a un nivell de confiança del 95% i contrast bilateral són:

$$F_{0,975(24,120)} \approx 1,76 \text{ i } F_{0,025(24,120)} \approx 0,50^1$$

⁽¹⁾Nota

$$F_{0,025(24,120)} = \frac{1}{F_{0,975(120,24)}} \approx \frac{1}{2,02} = 0,50$$

Com que el valor de l'estadístic de contrast obtingut (2,262) es troba fora de l'interval comprès entre els valors crítics 1,76 i 0,50, podem rebutjar la hipòtesi nul·la i podem concloure que, a partir de les nostres dades i amb un nivell de confiança del 95%, tenim evidència suficient per a determinar que el valor del coeficient alfa en la població no és de 0, i per tant aquest coeficient és estadísticament significatiu.

Per a construir l'interval confidencial d'aquest coeficient alfa, només cal substituir, en la fórmula de l'estadístic de contrast, els valors crítics de la distribució F i aïllar els valors de α :

$$\frac{1 - \alpha}{1 - 0,558} \leq 1,76$$

$$\alpha \geq 1 - 1,76(1 - 0,558) = 0,22$$

$$\frac{1 - \alpha}{1 - 0,558} \geq 0,50$$

$$\alpha \leq 1 - 0,50(1 - 0,558) = 0,78$$

Per tant, podem afirmar que amb un nivell de confiança del 95%, els valors del coeficient alfa en la població estaran compresos entre 0,22 i 0,78.

4. Contrast de dos coeficients en mostres independents

Per a saber si la diferència entre dos coeficients alfa obtinguts en mostres diferents de subjectes són iguals o no, hem d'aplicar un contrast per a dos coeficients en mostres independents. Per aplicar aquest contrast seguirem una sèrie de passos:

- Plantejament de les hipòtesis nul·la i alternativa

- Hipòtesi nul·la: $\hat{\alpha}_1 = \hat{\alpha}_2$

- Hipòtesi alternativa: $\hat{\alpha}_1 \neq \hat{\alpha}_2$

- Càlcul de l'estadístic de contrast:

$$w = \frac{1 - \hat{\alpha}_1}{1 - \hat{\alpha}_2} = \frac{1 - 0,558}{1 - 0,70} = 1,47$$

que es distribueix segons una F de Snedecor amb $(N_1 - 1)$ i $(N_2 - 1)$ graus de llibertat, essent N_1 i N_2 les grandàries de les dues mostres.

- Valors crítics de la distribució F de Snedecor amb 24 $(N_1 - 1)$ i 29 $(N_2 - 1)$ graus de llibertat, per a un nivell de confiança del 95% i contrast bilateral són:

$$F_{0,975(24,29)} = 2,15 \text{ i } F_{0,025(24,29)} = 0,45$$

Com que el valor de l'estadístic de contrast obtingut (1,47) cau dins de l'interval comprès entre els valors crítics (0,45 – 2,15), no tenim prou evidències per a rebutjar la hipòtesi nul·la, i per tant hem de concloure que la diferència entre els dos coeficients no és estadísticament significativa.

5. Contrast per a dos coeficients en mostres dependents

Com que hem administrat el mateix test dues vegades al mateix grup de subjectes, per a determinar si entre tots dos coeficients alfa obtinguts hi ha o no diferències estadísticament significatives, aplicarem un contrast per a dos coeficients en mostres dependents. Els passos a seguir són aquests:

- Hipòtesi nul·la: $\hat{\alpha}_1 = \hat{\alpha}_2$

- Hipòtesi alternativa: $\hat{\alpha}_1 \neq \hat{\alpha}_2$

- Càlcul de l'estadístic de contrast:

$$t = \frac{(\hat{\alpha}_1 - \hat{\alpha}_2)\sqrt{N-2}}{\sqrt{4(1-\hat{\alpha}_1)(1-\hat{\alpha}_2)(1-r_{12})}} = \frac{(0,558-0,64)\sqrt{25-2}}{\sqrt{4(1-0,558)(1-0,64)(1-0,92)}} = -1,74$$

que es distribueix segons una t de Student amb $N - 2$ graus de llibertat, essent N el nombre de subjectes de la mostra.

- Valors crítics de la distribució t de Student amb 23 ($N - 2$) graus de llibertat, per a un nivell de confiança del 95% i contrast bilateral són:

$$t_{0,975(23)} = 2,069 \text{ i } t_{0,025(23)} = -2,069$$

Com que l'estadístic de contrast obtingut ($-1,74$) queda dins de l'interval entre els valors crítics ($-2,069 - 2,069$), acceptem la hipòtesi nul·la i podem concloure que, amb un nivell de confiança del 95%, la diferència entre els dos coeficients alfa no és estadísticament significativa.

6. Càlcul del KR_{20} i el KR_{21} de Kuder-Richardson i comparació dels dos valors

La fórmula del KR_{20} simplement substitueix, en la del coeficient alfa de Cronbach, el sumatori de les variàncies dels ítems pel sumatori dels productes p_j per q_j , essent (p_j) la proporció de subjectes que encerten l'ítem o tenen un 1 en aquest ítem i (q_j) la proporció de subjectes que no l'encerten o tenen un 0. Així, per exemple en el primer ítem, hi ha nou subjectes amb una puntuació d'1, i setze amb una puntuació de 0. Per tant, p_j serà $9/25$ (0,36), i q_j $16/25$ (0,64).

La fórmula del KR_{20} és la següent:

$$KR_{20} = \frac{n}{n-1} \left(1 - \frac{\sum_{j=1}^n p_j q_j}{S_x^2} \right)$$

I els productes p_j per q_j seran:

	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6
p	0,36	0,56	0,68	0,24	0,68	0,72
q	0,64	0,44	0,32	0,76	0,32	0,28
$p \cdot q$	0,2304	0,2464	0,2176	0,1824	0,2176	0,2016

Per tant,

$$KR_{20} = \frac{n}{n-1} \left[1 - \frac{\sum_{j=1}^n p_j q_j}{S_x^2} \right] = \frac{6}{5} \left(1 - \frac{0,2304 + 0,2464 + 0,2176 + 0,1824 + 0,2176 + 0,2016}{2,4224} \right) = 0,558$$

El coeficient alfa calculat amb la fórmula del KR_{20} ens dona un valor de 0,558.

L'aplicació de la fórmula del KR_{21} per al nostre exercici ens dona el resultat següent:

$$KR_{21} = \frac{n}{n-1} \left(1 - \frac{\bar{X} - \bar{X}^2/n}{S_x^2} \right) = \frac{6}{5} \left(1 - \frac{3,24 - 3,24^2/6}{2,4224} \right) = 0,462$$

essent $\bar{X}$ la mitjana de les puntuacions totals dels vint-i-cinc subjectes de la mostra en el test. Aquesta mitjana és de 3,24.

Com podem observar, el valor del KR_{21} és inferior al del KR_{20} , essent aquest darrer igual al valor del coeficient alfa de Cronbach. Aquesta diferència és deguda al fet que no tots els ítems tenen la mateixa proporció d'uns i de zeros, o sigui que els valors de p_i i q_i no són iguals en tots els ítems i , per tant, l'aplicació de la fórmula KR_{21} per a aquest cas és inapropiada i no ens dóna un valor adequat per a estimar el coeficient alfa.

7. Per a determinar quina serà la nova fiabilitat del test si afegíssim tres ítems més, haurem d'aplicar la fórmula de Spearman Brown:

$$R_{xx} = \frac{k r_{xx}}{1 + (k-1)r_{xx}}$$

En què:

R_{xx} és el nou coeficient de fiabilitat del test allargat.

r_{xx} és el coeficient de fiabilitat del test original.

k és el nombre de vegades que s'allarga o s'escurça el test. D'aquesta manera k estarà determinat pel quocient entre el nombre d'ítems finals (n_f) del test dividit pel nombre d'ítems inicials (n_i) del test:

$$k = \frac{n_f}{n_i} = \frac{\text{Ítems finals}}{\text{Ítems inicials}} = \frac{9}{6} = 1,5$$

Així, la nova fiabilitat serà:

$$R_{xx} = \frac{k r_{xx}}{1 + (k-1)r_{xx}} = \frac{1,5 \times 0,558}{1 + (1,5-1)0,558} = 0,65$$

Per tant, si afegíssim tres ítems més al test, suposant que mesuressin el mateix constructe, la seva nova fiabilitat seria de 0,65.

8. Si volem saber quants ítems hauríem d'afegir al test inicial per a arribar a una fiabilitat de 0,70, hem d'aïllar k de la fórmula de Spearman-Brown:

$$k = \frac{R_{xx}(1-r_{xx})}{r_{xx}(1-R_{xx})} = \frac{0,70(1-0,558)}{0,558(1-0,70)} = 1,85$$

I, per tant, el nombre d'ítems que cal afegir serà $(k \cdot \text{Ítems inicials}) - \text{Ítems inicials} (1,85 \cdot 6) - 6 = 5,1$.

Si volem que el test tingui una fiabilitat de 0,70 hem d'afegir sis ítems als sis inicials.

9. Per a estimar la puntuació veritable que tindrà el primer subjecte de la matriu de dades del nostre exercici, que ha obtingut una puntuació total en el test de 3, amb un nivell de confiança del 95%, podem utilitzar dos procediments:

a) La distribució normal dels errors:

Els passos que s'han de seguir en aquest cas seran:

- Calcular l'error típic de mesura (S_e):

$$S_e = S_x \sqrt{1 - r_{xx}} = 1,5564 \sqrt{1 - 0,558} = 1,03$$

en què S_x és la desviació típica de les puntuacions del test i r_{xx} és el coeficient alfa obtingut.

- Cercar el valor $Z_{\alpha/2}$ que per al nivell de confiança del 95% segons les taules de la distribució normal és d'1,96.
- Calcular l'error màxim de mesura ($E_{m\grave{a}x}$):

$$E_{m\grave{a}x} = Z_{\alpha/2} \cdot S_e = 1,96 \cdot 1,03 \approx 2$$

- Calcular l'interval de confiança de la puntuació veritable del subjecte a partir de l'expressió següent:

$$IC = X \pm E_{m\grave{a}x} = 3 \pm 2$$

$$1 \leq V \leq 5$$

La puntuació veritable del primer subjecte de la matriu de dades, que ha obtingut una puntuació total de 3 en el test, amb un nivell de confiança del 95% i aplicant la distribució normal dels errors, estarà compresa entre 1 i 5 punts.

b) El model de la regressió:

Si utilitzem el model de la regressió per a obtenir aquest interval de confiança, seguirem aquests altres passos:

- Calculem la puntuació veritable estimada segons el model de la regressió:

$$V' = r_{xx}(X - \bar{X}) + \bar{X}$$

En què:

V' és la puntuació veritable pronosticada.

r_{xx} és el coeficient de fiabilitat del test.

X és la puntuació empírica obtinguda pel subjecte.

$\bar{X}$ és la mitjana de les puntuacions del test, que en el nostre cas és de 3,24.

$$V' = 0,558 \cdot (3 - 3,24) + 3,24 = 3,10$$

- Obtenim l'error típic d'estimació:

$$S_{VX} = S_X \sqrt{1 - r_{XX}} \sqrt{r_{XX}} = S_e \sqrt{r_{XX}} = 1,03 \sqrt{0,558} = 0,77$$

En què:

S_X és la desviació típica de les puntuacions del test.

r_{XX} és el coeficient de fiabilitat del test.

S_e és l'error típic de mesura.

- Calculem l'error màxim:

$$E_{màx} = 1,96 \cdot 0,77 = 1,51$$

- Obtenim l'interval de confiança:

$$V' \pm E_{màx}$$

$$3,10 \pm 1,51$$

$$1,59 \leq V \leq 4,61$$

Aplicant el model de la regressió, la puntuació veritable del primer subjecte de la matriu de dades, amb un nivell de confiança del 95%, estarà compresa entre 1,59 i 4,61.

7. Solucionari dels exercicis sobre validesa

Exercici 1

a) Per a obtenir el coeficient de validesa del test, hem de calcular el coeficient de correlació de Pearson entre les puntuacions dels subjectes en el test i en el criteri (BDI). Per tant, aplicarem la fórmula següent:

$$r_{xy} = \frac{n \sum xy - \sum x \sum y}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

En què x són les puntuacions dels subjectes en el test de sis ítems i y les seves puntuacions en el criteri (BDI).

Amb les nostres dades, aquest coeficient de fiabilitat és igual a 0,80.

De fet, en aquest cas també ens podríem plantejar utilitzar el coeficient de correlació de Spearman, ja que el nostre test, en tenir tan pocs ítems (només sis), es podria argumentar que la seva escala de mesura és ordinal i no d'interval o raó com requereix el coeficient de correlació de Pearson. El valor del coeficient de correlació de Spearman per a les nostres dades és de 0,82, que com veiem no és gaire diferent del de Pearson. De totes maneres, per als propers apartats continuarem operant amb el valor del coeficient de validesa de 0,80.

b) Per a obtenir el coeficient de validesa del test, suposant que aquest tingués una fiabilitat perfecta, aplicariem la fórmula d'atenuació per a aquest cas:

$$r_{v_{xy}} = \frac{r_{xy}}{\sqrt{r_{xx'}}}$$

Per tant, per al nostre exemple, el coeficient de validesa inicial del test és 0,80, i el seu coeficient de fiabilitat (calculat en els exercicis de fiabilitat amb el mètode de les dues meitats i la fórmula de Spearman-Brown) és de 0,707.

Per tant:

$$r_{v_{xy}} = \frac{r_{xy}}{\sqrt{r_{xx'}}} = \frac{0,80}{\sqrt{0,707}} = 0,95$$

El coeficient de validesa del test suposant-hi una validesa perfecta seria de 0,95.

c) Si afegim sis ítems més al test, en podrem calcular la nova fiabilitat amb l'expressió següent:

$$r_{Xy} = \frac{r_{xy}\sqrt{n}}{\sqrt{1+(n-1)r_{xx'}}}$$

En què:

r_{xy} és el valor inicial del coeficient de validesa del test.

$r_{xx'}$ és el coeficient de fiabilitat del test.

n és el nombre de cops que s'allarga el test.

Si allarguem el test amb sis ítems més del sis inicials, n serà igual a 2, ja que en dupliquem la longitud.

Per tant, el nou coeficient de validesa serà:

$$r_{Xy} = \frac{r_{xy}\sqrt{n}}{\sqrt{1+(n-1)r_{xx'}}} = \frac{0,80\sqrt{2}}{\sqrt{1+0,707}} = 0,84$$

d) En primer lloc, per a determinar la puntuació que podem pronosticar que tindrà en el criteri (BDI) un subjecte que ha obtingut 4 punts en el nostre test ($x = 4$), aplicarem el model de la regressió. L'equació de la recta de regressió serà:

$$y' = a + bx$$

En què b és el pendent de la recta i a la intersecció o ordenada en l'origen.

El pendent (b) i la intersecció (a) els podem calcular a partir de les fórmules següents:

$$b = r_{xy} \cdot \frac{s_y}{s_x} = 0,80 \times \frac{4,050}{1,556} = 2,08$$

$$a = \bar{y} - b \cdot \bar{x} = 12,56 - (3,24 \cdot 2,08) = 5,82$$

I, per tant, tenim que $y' = a + bx = 5,82 + (2,08 \times 4) = 14,14$.

Ara podem construir l'interval de confiança a partir de l'expressió següent:

$$IC_{1-\alpha} \rightarrow y' \pm t_{n-1;\alpha/2} \cdot s_{y-y'}$$

En què $t_{n-1,\alpha/2}$ és el valor de la t de Student amb $n - 1$ graus de llibertat i el nivell de confiança determinat, i $S_{y-y'}$ és l'error típic d'estimació.

En el nostre exercici, $t = 2,064$ i $s_{y-y'} = s_y \sqrt{1 - r_{xy}^2} = 4,050 \sqrt{1 - 0,80^2} = 2,43$.

Així,

$$IC_{1-\alpha} \rightarrow y' \pm t_{n-1,\alpha/2} \cdot s_{y-y'} = 14,14 \pm 2,064 \cdot 2,43 = [9,12 \quad 19,15]$$

Podem concloure que, a un subjecte que ha obtingut 4 punts en el nostre test, li podem pronosticar que tindrà entre 9 i 19 punts (arrodonint) en el BDI, amb un nivell de confiança del 95%.

e) Per a analitzar la validesa de decisió del nostre test, hem de construir la taula de contingència, a partir de la classificació dels subjectes entre els diagnosticats o no de trastorn depressiu lleu pel BDI i pel nostre test, en funció dels punts de tall de l'enunciat, en funció de les seves puntuacions directes del cas 2 en les dues proves.

Aquesta taula és la següent:

		BDI		Total
		No trastorn	Sí trastorn	
Test	No trastorn	14	1	15
	Sí trastorn	4	6	10
	Total	18	7	25

A partir d'aquesta taula podem obtenir:

- El percentatge d'acord entre els dos tests (P_c):

$$P_c = \frac{14+6}{25} = 0,80$$

Que podem interpretar com un nivell d'acord força elevat (80%).

- El coeficient kappa (k):

$$k = \frac{20 - 13,6}{25 - 13,6} = 0,56$$

Que podem interpretar com una relació acceptable entre les dues classificacions.

- La sensibilitat del nostre test:

$$\text{Sensibilitat} = \frac{\text{Diagnosticats pel nostre test}}{\text{Total diagnosticats (BDI)}} = \frac{6}{7} = 0,86$$

Que representa un 86% de diagnòstics correctes.

- L'especificitat del nostre test:

$$\text{Especificitat} = \frac{\text{No diagnosticats pel nostre test}}{\text{Total no diagnosticats (BDI)}} = \frac{14}{18} = 0,78$$

Que s'interpreta com un 76% d'encert en la detecció del no trastorn.

Exercici 2

Per obtenir les comunalitats dels diferents ítems en cada un dels dos components extrets, n'elevem al quadrat les saturacions factorials. La suma de les dues comunalitats en proporcionarà el valor de la comunalitat conjunta. El valor propi de cada component serà la suma de les comunalitats dels sis ítems, i la variància explicada per a cada un d'ells serà el percentatge que representa el valor propi respecte al total dels sis ítems. Així, tindrem la taula següent:

	Component		Comunalitats		
	1	2	C1	C2	Conjunta
Ítem 1	,587	,023	0,345	0,001	0,345
Ítem 2	,267	,793	0,071	0,629	0,700
Ítem 3	,085	,703	0,007	0,494	0,501
Ítem 4	,696	-,303	0,484	0,092	0,576
Ítem 5	,705	-,360	0,497	0,130	0,627
Ítem 6	,821	,219	0,674	0,048	0,722
Valor propi			2,079	1,393	3,472
Variància explicada			34,643	23,216	57,859

C1 és el component 1 i C2 el component 2.

La comunalitat de l'ítem 1 en el component 1 és igual a $0,587^2$ o sigui 0,345, i així per a la resta. La variància explicada pel primer component és igual al seu valor propi (2,079) dividit per 6 (nombre d'ítems) i multiplicat per 100:

$$\text{Variància explicada C1} = \frac{2,079}{6} \times 100 = 34,643.$$

La interpretació dels resultats d'aquest ACP aniria en la direcció de considerar el test amb una estructura interna bidimensional (mesuraria dues dimensions), ja que els diferents ítems presenten saturacions factorials elevades (superiors a 0,30) amb algun dels dos factors extrets. Així, els ítems 1, 4, 5 i 6

estarien relacionats amb el component 1, mentre que els ítems 2 i 3 estarien relacionats amb el component 2. El primer component explica el 34,643% de la variabilitat total dels sis ítems, mentre que el segon explica un percentatge del 23,216. Els dos components extrets conjuntament expliquen quasi un 58% de la variabilitat total.

8. Solucionari dels exercicis sobre transformació i interpretació de les puntuacions

Exercici 1

En primer lloc, per a fer qualsevol transformació de les puntuacions directes (X) hem de construir la taula de freqüències amb les freqüències absolutes (f_i) per a cada una de les puntuacions en el test. Aquestes freqüències absolutes seran el recompte del nombre de subjectes que han obtingut una puntuació determinada:

X	f_i
0	1
1	3
2	3
3	8
4	4
5	4
6	2

a) Percentils

Per a obtenir els percentils necessitem calcular les freqüències acumulades (f_a), els percentatges (P_i) i els percentatges acumulats (P_a):

X	f_i	f_a	P_i	P_a
0	1	1	4	4
1	3	4	12	16
2	3	7	12	28
3	8	15	32	60
4	4	19	16	76
5	4	23	16	92
6	2	25	8	100

A partir d'aquestes dades, per obtenir els percentils (P_c) apliquem la fórmula:

$$P_c = \frac{f_a + 0,5f_i}{N} \times 100$$

En què f_a és la freqüència acumulada prèvia a la puntuació directa de la qual es vol calcular el percentil, f_i la freqüència absoluta en la qual es troba la puntuació directa, i N el nombre de persones que constitueixen la mostra.

Així, per a la puntuació directa de 0, el percentil serà:

$$P_c = \frac{0 + (0,5 \cdot 1)}{25} \times 100 = 2$$

Per a la puntuació d'1:

$$P_c = \frac{1 + (0,5 \cdot 3)}{25} \times 100 = 10$$

Per a la puntuació de 2:

$$P_c = \frac{4 + (0,5 \cdot 3)}{25} \times 100 = 22$$

Per a la puntuació de 3:

$$P_c = \frac{7 + (0,5 \cdot 8)}{25} \times 100 = 44$$

Per a la puntuació de 4:

$$P_c = \frac{15 + (0,5 \cdot 4)}{25} \times 100 = 68$$

Per a la puntuació de 5:

$$P_c = \frac{19 + (0,5 \cdot 4)}{25} \times 100 = 84$$

Per a la puntuació de 6:

$$P_c = \frac{23 + (0,5 \cdot 2)}{25} \times 100 = 96$$

x	f_i	P_c
0	1	2
1	3	10
2	3	22
3	8	44

X	f_i	P_c
4	4	68
5	4	84
6	2	96

b) Puntuacions estandarditzades

La fórmula de les puntuacions estandarditzades és:

$$Z_x = \frac{X - \bar{X}}{S_x}$$

En què X és la puntuació directa, $\bar{X}$ la mitjana de la mostra, i S_x la desviació típica. Per tant, hem d'obtenir la mitjana ($\bar{X} = 3,24$) i la desviació típica ($S_x = 1,56$) de la distribució de dades, i aplicar la fórmula per a cada puntuació directa.

Per exemple, per a la puntuació de 0, la puntuació estandarditzada serà:

$$Z_x = \frac{X - \bar{X}}{S_x} = \frac{0 - 3,24}{1,56} = -2,08$$

Aplicant la fórmula a totes les puntuacions obtindrem:

X	Z_x
0	-2,08
1	-1,44
2	-0,79
3	-0,15
4	0,49
5	1,13
6	1,77

c) Puntuacions T

A partir de les puntuacions estandarditzades podem obtenir fàcilment les puntuacions T de McCall a partir de la transformació següent:

$$T = 50 + 10 Z_x$$

En el nostre cas, per a una puntuació directa de 0, la puntuació T serà $50 + (10 \times -2,08) = 29,2$, que arrodonim a l'enter més proper (29), ja que les puntuacions T no tenen decimals.

Per a les altres puntuacions tindrem:

X	Z_x	T
0	-2,08	29
1	-1,44	36
2	-0,79	42
3	-0,15	48
4	0,49	55
5	1,13	61
6	1,77	68

d) Enneatipus (E)

Per als enneatipus, primer hem de calcular les puntuacions estandarditzades normalitzades (Z_n), transformant els percentils en les puntuacions estandarditzades que sota la taula de la distribució normal tenen associada una proporció igual al percentil dividit per 100. Per exemple, al percentil 2, que en les nostres dades correspon a la puntuació directa de 0, correspon una puntuació estandarditzada normalitzada de -2,05, que és la puntuació z de la distribució normal associada a una proporció de 0,02 (percentil dividit per 100).

Un cop tenim la Z_n , per obtenir l'enneatipus simplement apliquem la transformació $E = 5 + 2 Z_n$, que per a la puntuació de 0 ens donarà $E = 5 + (2 - 2,05) = 0,9$, que arrodonim a l'enter més proper (1), ja que els enneatipus tampoc tenen decimals.

Per a les altres puntuacions obtindrem:

X	P_c	Z_n	E
0	2	-2,05	1
1	10	-1,28	2
2	22	-0,77	3
3	44	-0,15	5
4	68	0,47	6
5	84	0,99	7

X	P_c	Z_n	E
6	96	1,75	9

e) Decatipus (D)

Els decatipus també són puntuacions normalitzades derivades que s'obtenen de les puntuacions estandarditzades normalitzades a partir de la transformació següent: $D = 5,5 + 2 Z_n$, que per la puntuació de 0 ens donarà $D = 5,5 + (2 - 2,05) = 1,4$, que arrodonim a l'enter més proper (1) per la mateixa raó de l'apartat anterior.

Per a les altres puntuacions obtindrem:

X	P_c	Z_n	D
0	2	-2,05	1
1	10	-1,28	3
2	22	-0,77	4
3	44	-0,15	5
4	68	0,47	6
5	84	0,99	7
6	96	1,75	9

Exercici 2

Per calcular els percentils corresponents a les puntuacions directes, seguim els passos que veurem a continuació.

En primer lloc, obtenim la puntuació directa dels tres primers subjectes de la matriu de dades sumant el nombre d'ítems que cada un d'ells contesta correctament.

Així, per al primer subjecte la puntuació directa és de 9, per al segon de 6, i per al tercer de 2.

En segon lloc, obtenim la taula de freqüències absolutes (f_i) i de freqüències acumulades (f_a) de totes les puntuacions (X) obtingudes pels vint-i-cinc subjectes, fent el recompte del nombre de subjectes que obtenen cada una de les puntuacions.

X	f_i	f_a
1	2	2

X	f_i	f_a
2	4	6
3	2	8
4	2	10
5	2	12
6	3	15
7	4	19
8	3	22
9	3	25

En tercer lloc, apliquem la fórmula dels percentils per a les puntuacions dels tres primers subjectes:

- Primer subjecte:

$$X = 9 \quad P_c = \frac{f_a + 0,5f_i}{N} \times 100 = \frac{22 + (0,5 \cdot 3)}{25} \times 100 = 94$$

- Segon subjecte:

$$X = 6 \quad P_c = \frac{f_a + 0,5f_i}{N} \times 100 = \frac{12 + (0,5 \cdot 3)}{25} \times 100 = 54$$

- Tercer subjecte:

$$X = 2 \quad P_c = \frac{f_a + 0,5f_i}{N} \times 100 = \frac{2 + (0,5 \cdot 4)}{25} \times 100 = 16$$

D'altra banda, per a transformar les puntuacions directes en QI, hem d'obtenir, en primer lloc, la puntuació estandarditzada z i, en segon lloc, transformar aquesta puntuació z en QI.

La fórmula de la puntuació z és:

$$Z_x = \frac{X - \bar{X}}{S_x}$$

I requereix el càlcul previ de la mitjana i la desviació típica. En aquest exemple, la mitjana de les puntuacions dels vint-i-cinc subjectes és de 5,24, i la desviació típica de 2,61.

Així, les puntuacions estandarditzades dels tres primers subjectes seran:

- Primer subjecte:

$$X = 9 \quad Z_x = \frac{X - \bar{X}}{S_x} = \frac{9 - 5,24}{2,61} = 1,44$$

- Segon subjecte:

$$X = 6 \quad Z_x = \frac{X - \bar{X}}{S_x} = \frac{6 - 5,24}{2,61} = 0,29$$

- Tercer subjecte:

$$X = 2 \quad Z_x = \frac{X - \bar{X}}{S_x} = \frac{2 - 5,24}{2,61} = -1,24$$

Amb aquestes puntuacions ja podem obtenir els valors dels quocients intel·lectuals (QI) dels tres subjectes. La transformació en QI segueix aquesta expressió:

$$QI = 100 + (z \times 15)$$

- Primer subjecte: $QI = 100 + (1,44 \times 15) \approx 122$ (arrodonint a l'enter més proper)
- Segon subjecte: $QI = 100 + (0,29 \times 15) \approx 104$
- Tercer subjecte: $QI = 100 + (-1,24 \times 15) \approx 81$

Podem resumir aquestes transformacions en la taula següent:

Subjecte	Puntuació directa	Percentil	QI
1	9	94	122
2	6	54	104
3	2	16	81

9. Solucionari dels exercicis sobre anàlisi dels ítems

Exercici 1

Per a obtenir els índexs de dificultat (ID) i de discriminació (IDr) de cada ítem podem construir les taules següents, que resumeixen les respostes dels vint-i-cinc subjectes a cada un d'aquests tres ítems:

Ítem 1:

	Alternatives de resposta			
Grup de rendiment	A*	B	C	D
Alt	4	1	1	0
Mitjà	1	4	4	4
Baix	0	3	2	1

* Alternativa correcta

- Índex de dificultat:

$$ID = \frac{A - \frac{E}{K-1}}{N} = \frac{5 - \frac{20}{3}}{25} = -0,07$$

En què A és el nombre de subjectes que encerten l'ítem (en l'ítem 1 són 5), E el nombre de subjectes que no l'encerten (20), K el nombre d'alternatives de resposta de l'ítem (4), i N el nombre total de subjectes que responen a l'ítem (25).

- Índex de discriminació:

$$D = P_a - P_b = \frac{4}{6} - \frac{0}{6} = 0,67$$

En què P_a és la proporció de subjectes del grup d'alt rendiment que encerten l'ítem (per a l'ítem 1 són 4 de 6), i P_b és la proporció de subjectes del grup de baix rendiment que encerten l'ítem (0 de 6).

En funció d'aquests valors, podem considerar que l'ítem 1 és un ítem amb una dificultat molt elevada i amb una alta capacitat de discriminació.

Ítem 2:

	Alternatives de resposta			
Grup de rendiment	A	B	C*	D
Alt	0	0	6	0
Mitjà	2	0	11	0
Baix	0	1	3	2

* Alternativa correcta

- Índex de dificultat:

$$ID = \frac{A - \frac{E}{K-1}}{N} = \frac{20 - \frac{5}{3}}{25} = 0,73$$

- Índex de discriminació:

$$D = Pa - Pb = \frac{6}{6} - \frac{3}{6} = 0,50$$

Interpreten aquests valors en el sentit de considerar l'ítem 2 com un ítem fàcil i amb una alta discriminació.

Ítem 3:

	Alternatives de resposta			
Grup de rendiment	A	B	C	D*
Alt	0	0	0	6
Mitjà	6	0	3	4
Baix	2	1	3	0

* Alternativa correcta

- Índex de dificultat:

$$ID = \frac{A - \frac{E}{K-1}}{N} = \frac{10 - \frac{15}{3}}{25} = 0,20$$

- Índex de discriminació:

$$D = Pa - Pb = \frac{6}{6} - \frac{0}{6} = 1$$

Podem interpretar aquests valors en el sentit de considerar l'ítem 3 com un ítem força difícil i amb una altíssima capacitat de discriminació.

Exercici 2

Per a fer l'anàlisi dels distractors de cada un dels tres primers ítems, podem aprofitar la taula confeccionada en l'exercici anterior i calcular els índexs de discriminació de cada una de les alternatives de respostes incorrectes:

Ítem 1:

	Alternatives de resposta			
Grup de rendiment	A*	B	C	D
Alt	4	1	1	0
Mitjà	1	4	4	4
Baix	0	3	2	1

* Alternativa correcta

- Índex de discriminació de l'alternativa B:

$$D = P_a - P_b = \frac{1}{6} - \frac{3}{6} = -0,33$$

- Índex de discriminació de l'alternativa C:

$$D = P_a - P_b = \frac{1}{6} - \frac{2}{6} = -0,17$$

- Índex de discriminació de l'alternativa D:

$$D = P_a - P_b = \frac{0}{6} - \frac{1}{6} = -0,17$$

Com podem observar, els tres índexs són negatius, la qual cosa s'interpreta com que les tres alternatives de resposta presenten una discriminació adequada com a distractors. També cal comentar que l'ítem 1 és un ítem força difícil, i en conseqüència hi ha una proporció elevada de subjectes que escullen respostes incorrectes. Observant aquestes respostes incorrectes, també podem concloure que les tres alternatives són adequades, ja que no hi ha grans diferències quant al nombre de subjectes que escullen cada alternativa. Així, l'alternativa B és escollida per vuit dels vint-i-cinc subjectes, la C per set, i la D per cinc.

Ítem 2:

	Alternatives de resposta			
Grup de rendiment	A	B	C*	D
Alt	0	0	6	0
Mitjà	2	0	11	0

* Alternativa correcta

	Alternatives de resposta			
Baix	0	1	3	2

* Alternativa correcta

- Índex de discriminació de l'alternativa A:

$$D = P_a - P_b = \frac{0}{6} - \frac{0}{6} = 0$$

- Índex de discriminació de l'alternativa B:

$$D = P_a - P_b = \frac{0}{6} - \frac{1}{6} = -0,17$$

- Índex de discriminació de l'alternativa D:

$$D = P_a - P_b = \frac{0}{6} - \frac{2}{6} = -0,33$$

En aquest cas, dos dels índexs de discriminació són negatius, i l'altre és igual a 0. Podem interpretar que les alternatives de resposta B i D presenten una discriminació adequada com a distractors, però l'alternativa A, no (atès que una puntuació de 0 indica que ambdós grups mostren un rendiment similar). De totes maneres, cal considerar que l'ítem té una dificultat molt baixa i, per tant, les alternatives incorrectes són escollides per molt pocs subjectes.

Ítem 3:

	Alternatives de resposta			
Grup de rendiment	A	B	C	D*
Alt	0	0	0	6
Mitjà	6	0	3	4
Baix	2	1	3	0

* Alternativa correcta

- Índex de discriminació de l'alternativa A:

$$D = P_a - P_b = \frac{0}{6} - \frac{2}{6} = -0,33$$

- Índex de discriminació de l'alternativa B:

$$D = P_a - P_b = \frac{0}{6} - \frac{1}{6} = -0,17$$

- Índex de discriminació de l'alternativa C:

$$D = Pa - Pb = \frac{0}{6} - \frac{3}{6} = -0,50$$

Els tres índexs de discriminació són adequats pel que fa a les característiques requerides als distractors. De totes maneres, cal considerar que l'alternativa *B* s'hauria de revisar, ja que només és escollida per un subjecte *i*, per tant, podem considerar que és massa evident que no és correcta.

Activitats pràctiques. Solucions utilitzant el programa Excel

PID_00241081

Maite Barrios
Antoni Cosculluela

Temps mínim de dedicació recomanat: 9 hores



**Maite Barrios**

Professora agregada del Departament de Metodologia de les Ciències del Comportament de la Universitat de Barcelona (UB) i consultora dels Estudis de Psicologia i Ciències de l'Educació de la Universitat Oberta de Catalunya (UOC). Doctora en Psicologia per la UB. Desenvolupa la docència en l'àmbit de la metodologia impartint, entre d'altres, les assignatures de Psicometria i Estadística aplicada a la psicologia i a informació i documentació. Coautora de diversos llibres i textos sobre estadística i tècniques de recerca. La seva recerca se centra, principalment, en la infometria, el biaix de gènere en ciència i els estudis bibliomètrics.

**Antoni Cosculluela**

Professor titular d'universitat del Departament de Metodologia de les Ciències del Comportament de la Facultat de Psicologia de la Universitat de Barcelona (UB) i consultor dels Estudis de Psicologia i Ciències de l'Educació de la Universitat Oberta de Catalunya (UOC). Doctor en Psicologia per la UB. La seva tasca docent s'ha desenvolupat en l'àmbit de la metodologia i els dissenys de recerca, i de l'anàlisi de dades en psicologia i en informació i documentació. Autor i coautor de diferents llibres d'estadística aplicada. Actualment la seva recerca s'orienta, entre altres camps, als estudis bibliomètrics en diferents àrees de coneixement.

Índex

Introducció.....	5
1. Introducció als aspectes bàsics del programa Excel.....	7
1.1. Com dissenyar una funció?	8
1.2. Què són les funcions preconfigurades?	10
1.3. Activació i accés dels programes d'anàlisi preconfigurades	12
2. Solucions dels exercicis de fiabilitat.....	16
2.1. Exercici 1	16
2.2. Exercici 2	26
3. Solucions dels exercicis de validesa.....	52
3.1. Exercici 1	52
3.2. Exercici 2	68
4. Solucions dels exercicis de transformació i interpretació de puntuacions.....	72
4.1. Exercici 1	72
4.2. Exercici 2	86
5. Solucions dels exercicis d'anàlisi d'ítems.....	98
5.1. Exercici 1	98
5.2. Exercici 2	110

Introducció

El document que presentem complementa el contingut del mòdul d'«Activitats pràctiques». El seu propòsit és servir de guia a l'estudiant a l'hora de realitzar les activitats pràctiques proposades en el citat mòdul utilitzant el programa Excel.

La versió d'Excel que s'ha utilitzat en aquest material es la versió Microsoft Excel 2010. Malgrat la contínua actualització del programari de Microsoft, la majoria de comandaments i menús als quals fem referència en aquest material són similars tant en les versions prèvies com posteriors.

L'Excel, igual que molts dels programes de Microsoft, es caracteritza per disposar de diferents vies per fer una mateixa acció. Activar un menú o seleccionar una opció concreta pot realitzar-se de diverses maneres. En aquest material presentem la via més comuna o senzilla per realitzar les diferents accions, però això no vol dir que sigui ni l'única ni la més ràpida. Vosaltres mateixos podeu provar vies alternatives per realitzar les mateixes accions que anirem descrivint al llarg del document.

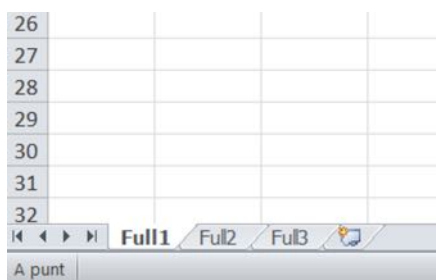
El material està organitzat en cinc apartats. Un primer apartat pretén ser una breu introducció als aspectes més bàsics del programa Excel, i en especial a tots aquells aspectes que tenen relació amb l'anàlisi de dades. Els apartats següents se centren en la resolució de les activitats pràctiques del mòdul d'«Activitats pràctiques» del material de *Psicometria* a partir del programa Excel. Per aquest ordre es resoldran pas a pas els exercicis d'aquest mòdul que fan referència als apartats de «Fiabilitat», «Validesa», «Transformació i interpretacions de puntuacions» i «Anàlisi dels ítems».

1. Introducció als aspectes bàsics del programa Excel

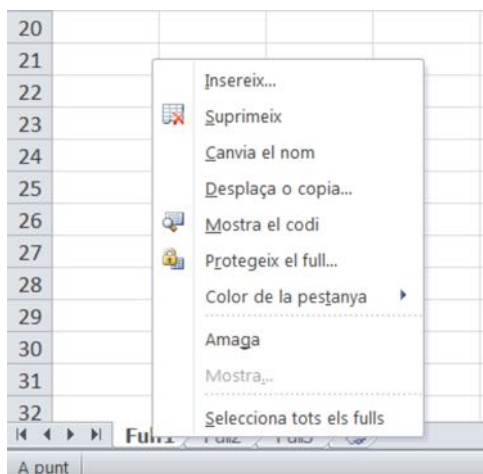
L'Excel forma part del paquet de programes de Microsoft Office i es caracteritza per ser un programa del tipus **full de càlcul**, la qual cosa permet realitzar operacions amb valors numèrics organitzats en una quadrícula.

El programa presenta menús similars a molts altres programes de Microsoft Office, però cal destacar algunes característiques diferenciadores, com per exemple la seva organització en **fileres** i **columnes**. Les fileres s'identifiquen per números, mentre que les columnes s'identifiquen per lletres. D'aquesta manera podem identificar cada casella de la quadrícula per una lletra i per un número.

Una altra de les característiques de l'Excel és que s'articula com un **llibre** format per diferents fulls independents (per defecte n'apareixen tres). A la part inferior esquerra de la pantalla del programa veureu indicats aquests fulls.



És molt fàcil moure's d'un full a l'altre, només cal prémer amb el ratolí a sobre de cada un dels fulls. A més, utilitzant el botó dret del ratolí s'obre un menú que us permetrà operar amb els fulls. Per exemple, entre altres funcions, us permetrà afegir un nou full, eliminar-ho, canviar-li el nom, moure'l de lloc, etc.



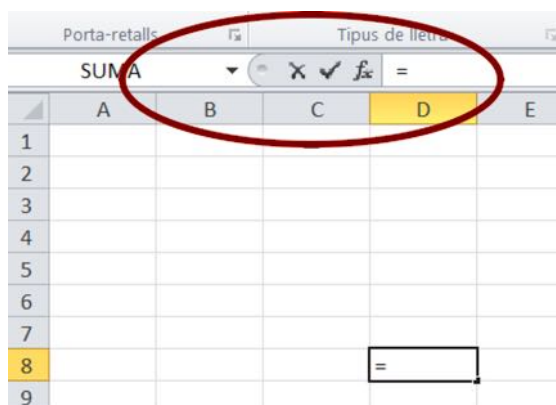
Bàsicament l'Excel disposa de tres formes per operar amb números:

- A través del disseny de funcions.
- A través de funcions preconfigurades.
- A partir de programes d'anàlisi preconfigurats.

A continuació, veurem com operar amb cada una d'elles utilitzant diferents exemples.

1.1. Com dissenyar una funció?

Per crear una funció cal que us situeu a una casella buida i escriviu el signe igual (=). A la barra de fórmules apareixerà el mateix que escriviu a la casella:



A partir d'aquí podeu operar com si l'Excel es tractés d'una calculadora. Per exemple, podeu representar expressions aritmètiques:

SUMA X ✓ f _x =429+58				
	A	B	C	D
1				
2				
3				=429+58
4				
5				

O bé fer referència a caselles que contenen determinats valors:

SUMA X ✓ f _x =B3+C3				
	A	B	C	D
1				
2				
3		429	58	=B3+C3
4				
5				

Una fórmula es pot copiar a altres caselles a partir de l'ordre d'autoemplenar. En aquest cas, caldrà que us situeu a la banda inferior dreta de la casella i premeu dues vegades la creu en negreta que apareix. Per exemple, en la figura següent s'observen unes dades d'un test de quatre ítems que han contestat cinc subjectes. En aquest cas ens interessa calcular la puntuació total d'un test. Així, una vegada hem calculat la suma del primer subjecte, caldrà prémer dues vegades la creu que apareix a la part inferior dreta de la casella i el programa emplenarà la resta de caselles amb la puntuació total del test per a la resta de subjectes (d'ara endavant anomenarem aquesta acció **autoemplenament**).

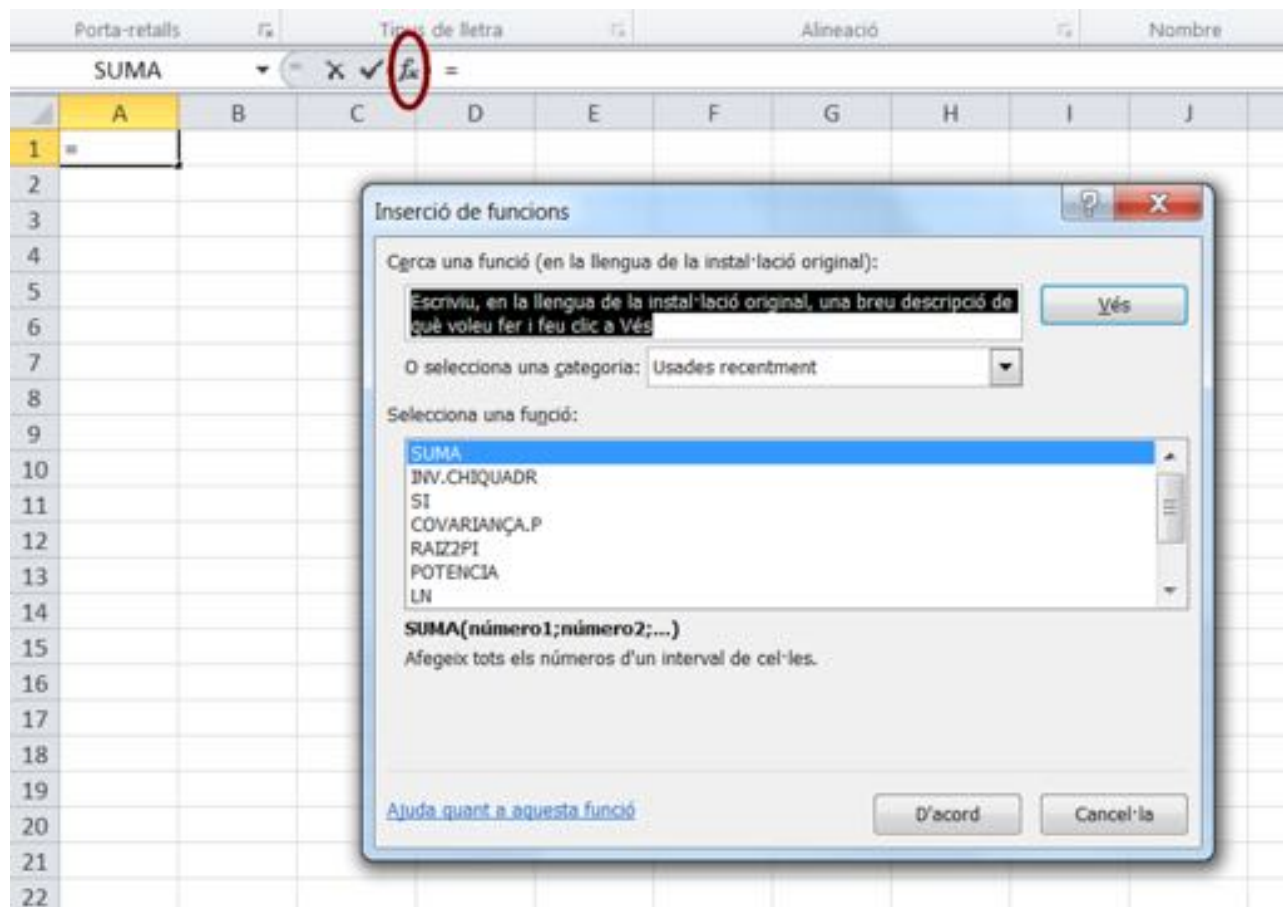
F2							f_x	=B2+C2+D2+E2
	A	B	C	D	E	F		
1	Subjecte	item 1	item 2	item 3	item 4	Total		
2		1	1	0	1	1	3	
3		2	0	0	0	0		
4		3	1	0	1	1		
5		4	1	1	1	1		
6		5	1	1	0	1		
7								
8								

F2							f_x	=B2+C2+D2+E2
	A	B	C	D	E	F		
1	Subjecte	item 1	item 2	item 3	item 4	Total		
2		1	1	0	1	1	3	
3		2	0	0	0	0	0	
4		3	1	0	1	1	3	
5		4	1	1	1	1	4	
6		5	1	1	0	1	3	
7								

1.2. Què són les funcions preconfigurades?

Les **funcions preconfigurades** són funcions d'ús comú que estan implementades en l'Excel.

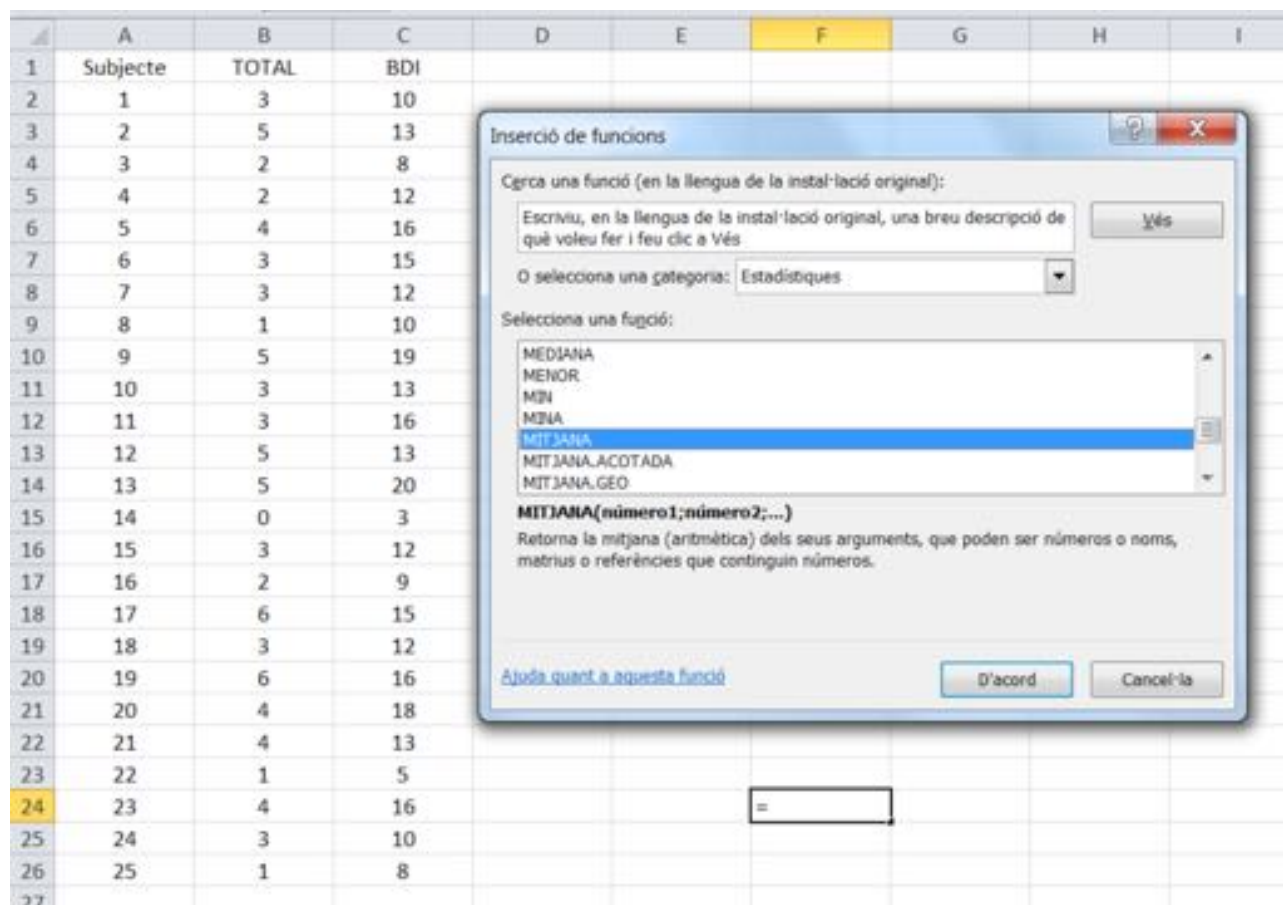
Podeu activar-les si premeu la icona d'**Inserció de funcions** a la barra d'eines. Observareu que apareix un nou menú en el qual podreu seleccionar aquella funció que més us interessi.



Per exemple, us pot interessar calcular la mitjana de la puntuació total del test de depressió de la matriu de dades descrita en el cas 2 de les activitats pràctiques. Per calcular la puntuació mitjana caldrà que seguiu els passos següents:

- 1) Situeu-vos en una casella buida (el resultat us apareixerà en la casella on us situeu).
- 2) Prémer la icona d'**Inserció de funcions**.
- 3) Seleccioneu la categoria **Estadístiques**.
- 4) Seleccioneu la funció **MITJANA**.

5) Premeu D'acord.

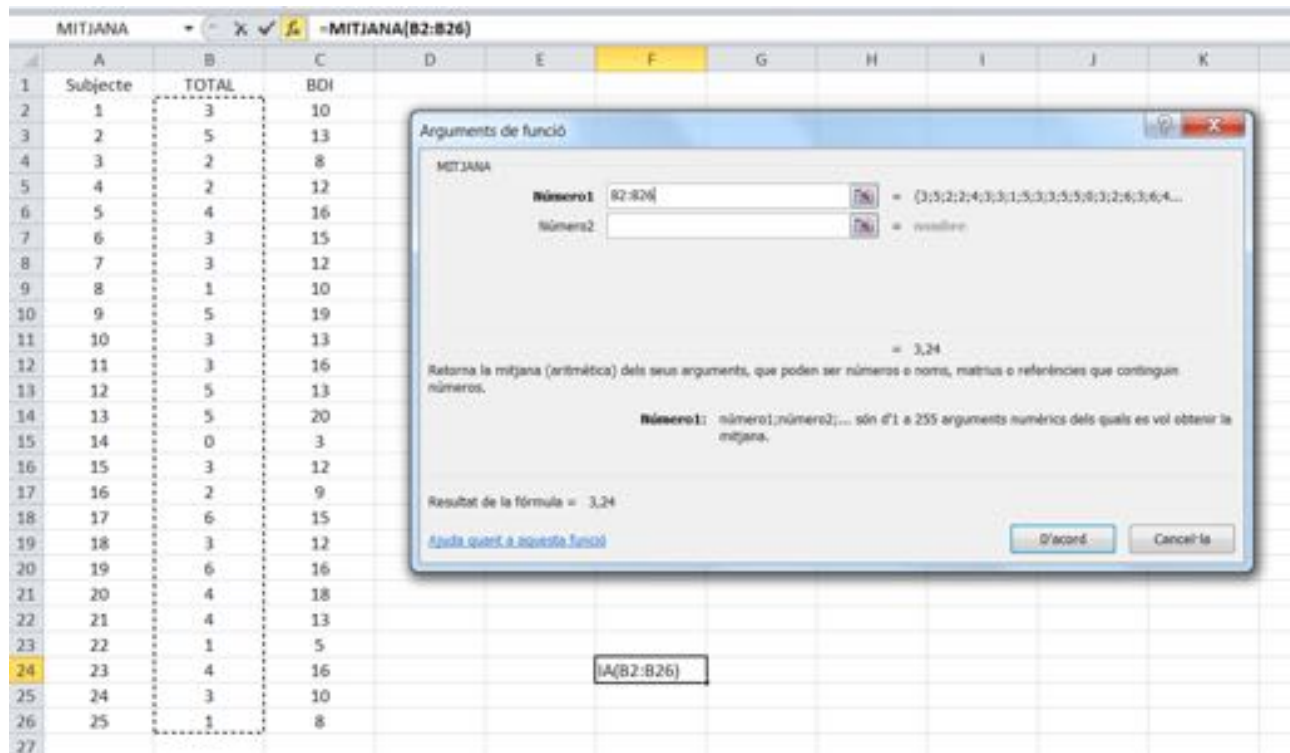


6) En la nova pantalla que us apareix, situeu-vos en el camp «Número 1».

7) Seleccioneu el rang corresponent (utilitzant el ratolí) o bé especifiqueu les caselles on estan situades les dades (en el nostre cas «B1:B26»).

8) Premeu D'acord.

Tal com hem comentat, el resultat us apareixerà a la casella on us heu situat.



Algunes de les funcions preconfigurades que utilitzarem durant l'assignatura seran les següents:

Categoria	
Matemàtiques i trigonomètriques	Estadístiques
Suma	Coefficient de correlació
Arrel	Desviació típica
Producte	Freqüència
Potència	Mitjana

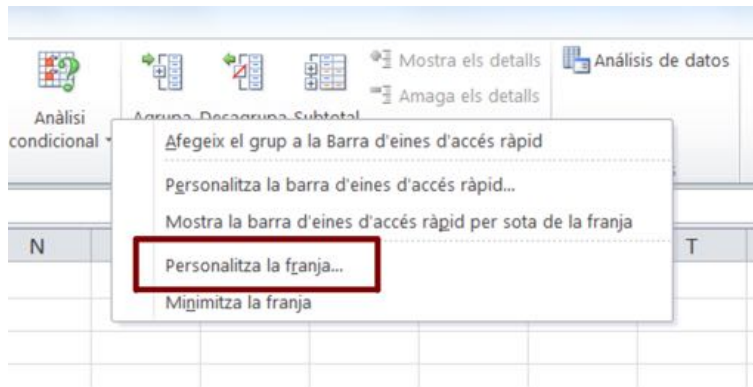
1.3. Activació i accés dels programes d'anàlisi preconfigurades

L'Excel inclou en el bloc d'Anàlisis del menú de Dades l'opció d'Anàlisis de datos, la qual us permet fer alguns càlculs estadístics.



Aquesta opció no està activada per defecte, per la qual cosa el primer que caldrà fer serà activar-la seguint els passos següents:

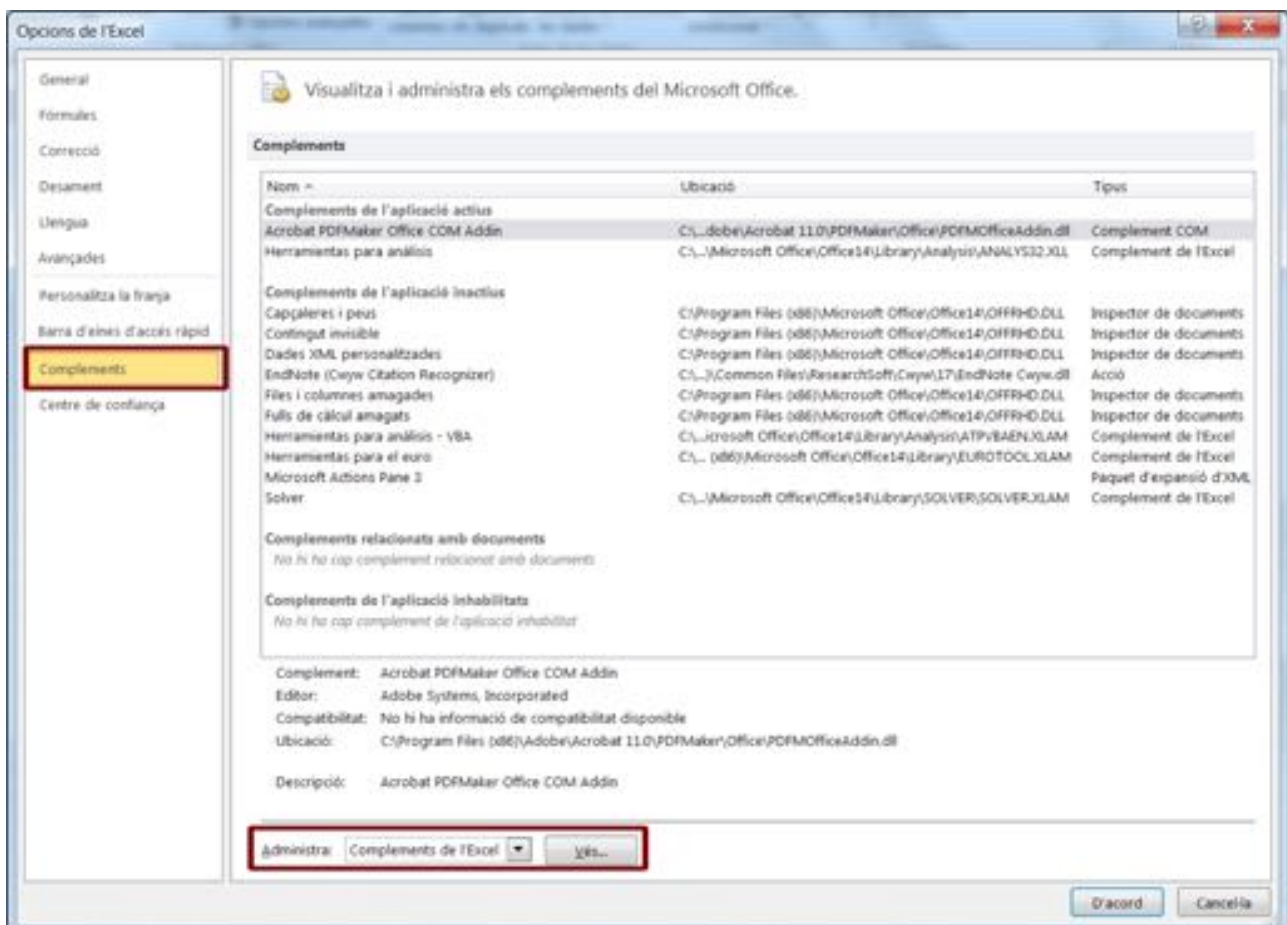
- 1) Premeu amb el botó dret del ratolí a sobre del submenú d'**Anàlisi**.
- 2) Seleccioneu l'opció **Personalitza la franja...**



Nota

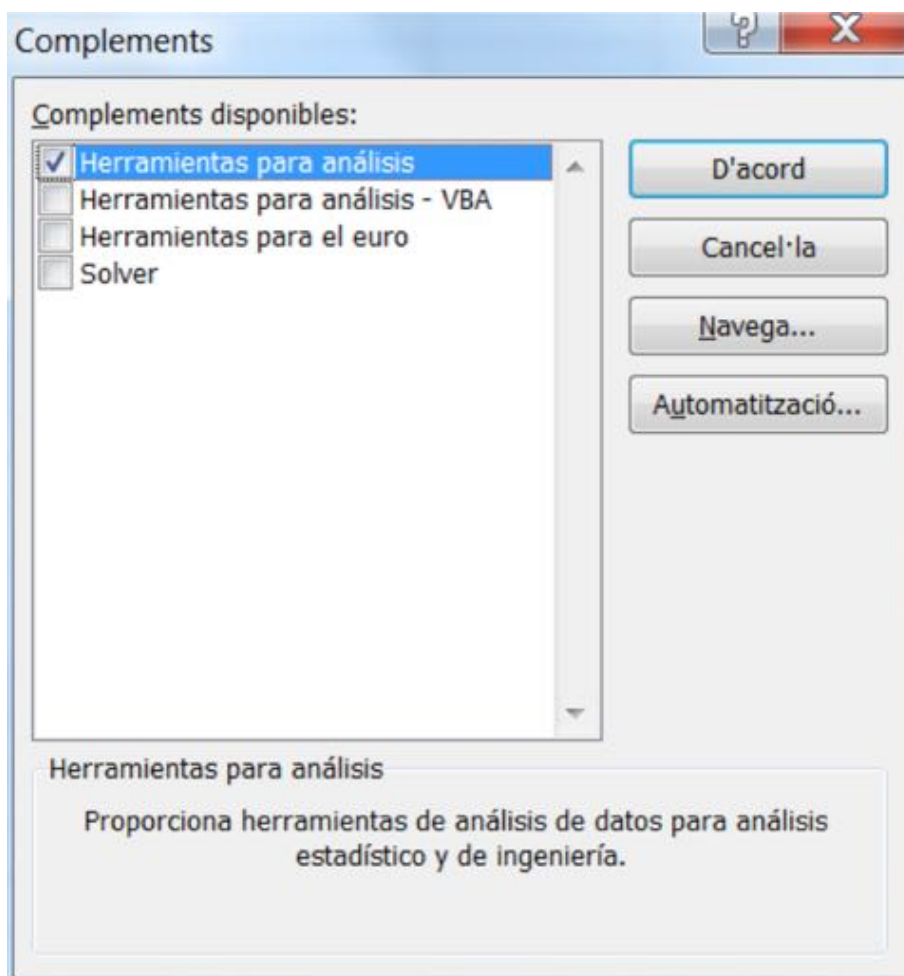
Cal tenir en compte que en funció del sistema operatiu i de la versió d'Excel que tingueu instal·lada aquestes instruccions poden variar. En aquest cas es recomana consultar l'ajuda del programa o la pàgina oficial de Microsoft Excel.

- 3) En la nova pantalla que apareix seleccioneu **Complements**.
- 4) Premeu el botó **Ves** en **Administra: Complements d'Excel**.

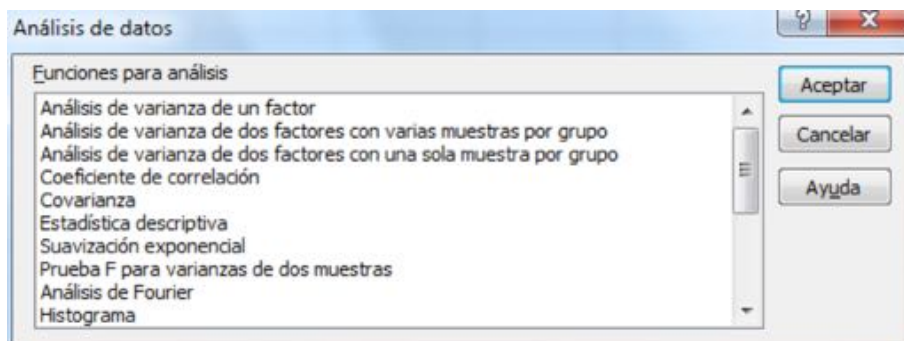


5) En la nova pantalla de **Complements**, seleccioneu **Herramientas para análisis**.

6) Premeu **D'acord** i us apareixerà en el menú d'**Análisis** l'opció d'**Análisis de datos**.



Si premeu a sobre de l'opció **Análisis de datos** s'obrirà la pantalla següent:



Algunes de les funcions d'anàlisi que us seran útils al llarg de l'assignatura i que anirem descrivint al llarg d'aquest document són:

- Coeficient de correlació.

- Estadística descriptiva.
- Regressió.
- Prova t.

2. Solucions dels exercicis de fiabilitat

2.1. Exercici 1

A partir de l'enunciat del cas 1, resoleu les qüestions següents:

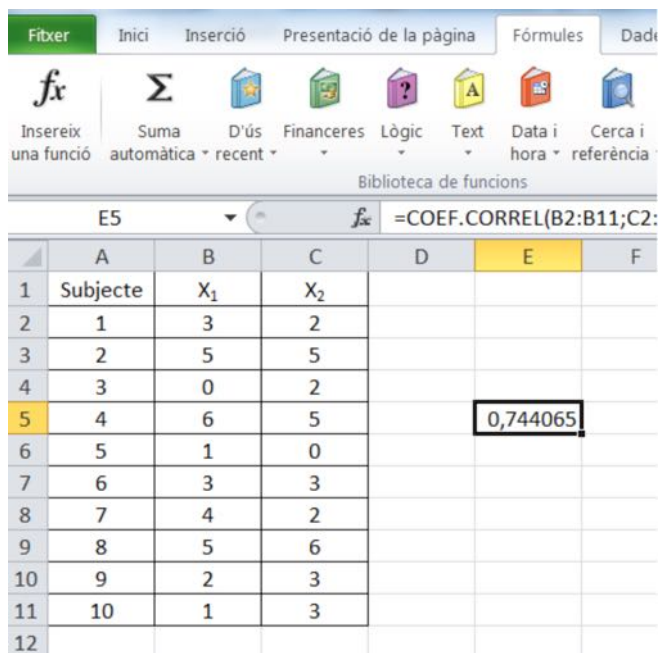
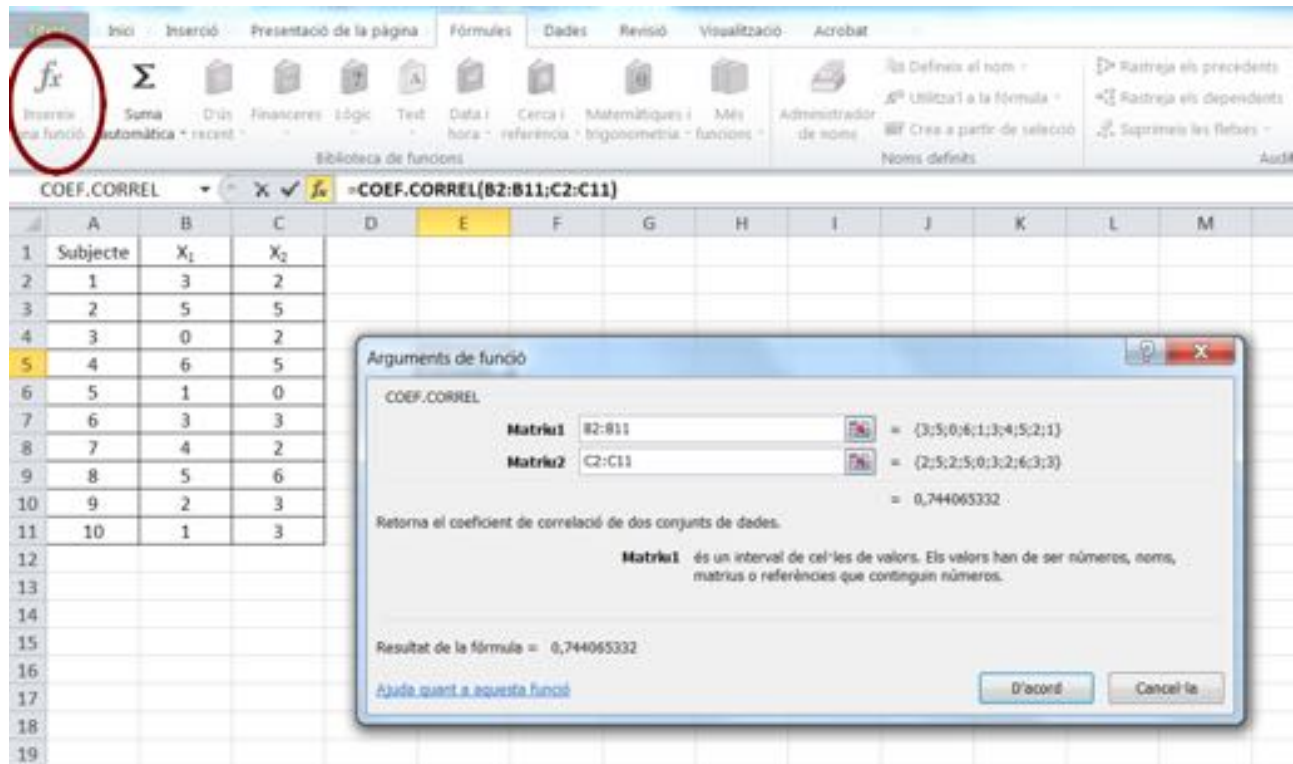
1. Obteniu el coeficient de fiabilitat del test.

En aquest cas, com tenim les puntuacions d'una mostra de subjectes en les dues formes paral·leles d'un test, obtindrem el seu coeficient de fiabilitat aplicant el mètode de les formes paral·leles, i per tant calcularem el coeficient de correlació de Pearson entre les puntuacions dels subjectes en aquestes dues formes.

Per obtenir un coeficient de correlació de Pearson amb l'Excel, podeu fer-ho amb la funció **COEF.DE.CORREL** o amb el programa preconfigurat corresponent.

Amb la funció:

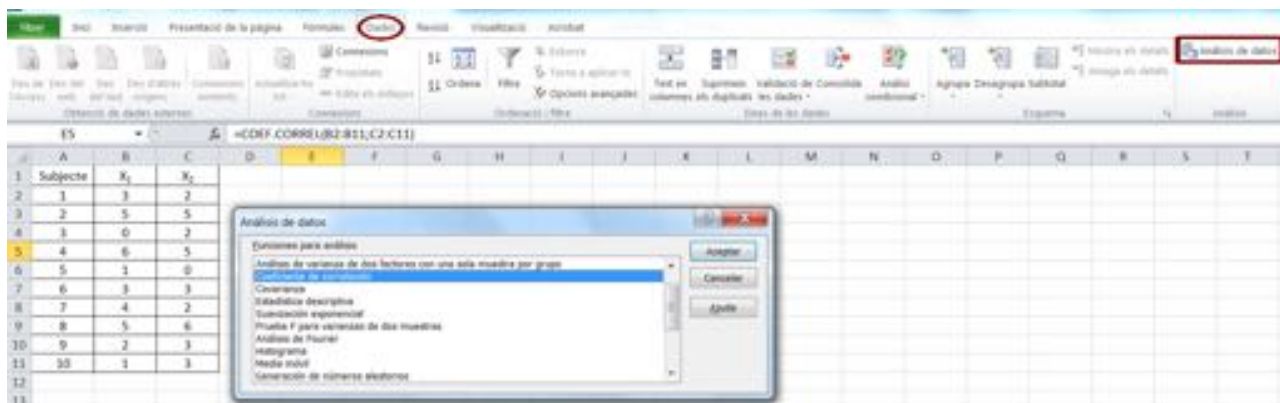
- 1) Inserir les dades del Cas 1 en un full.
- 2) Col·loqueu el cursor en una casella buida (per exemple, l'E3).
- 3) Seguiu la seqüència: **Fórmules, Insereix una funció**.
- 4) Seleccioneu dins la categoria **Estadístiques** la funció **COEF.DE.CORREL**.
- 5) Introduïu a **Matriu1** el rang de les caselles on teniu les dades de x_1 «B2:B11» i a **Matriu2** el rang de les caselles on teniu les dades de x_2 «C2:C11».
- 6) Premeu **D'acord**.



En la casella E3 teniu el resultat del coeficient de correlació de Pearson, que és de 0,744.

Amb el programa preconfigurat:

1) Un cop heu activat el complement d'Herramientas para análisis, seguiu la seqüència Dades, Análisis de datos, Coeficiente de correlación.

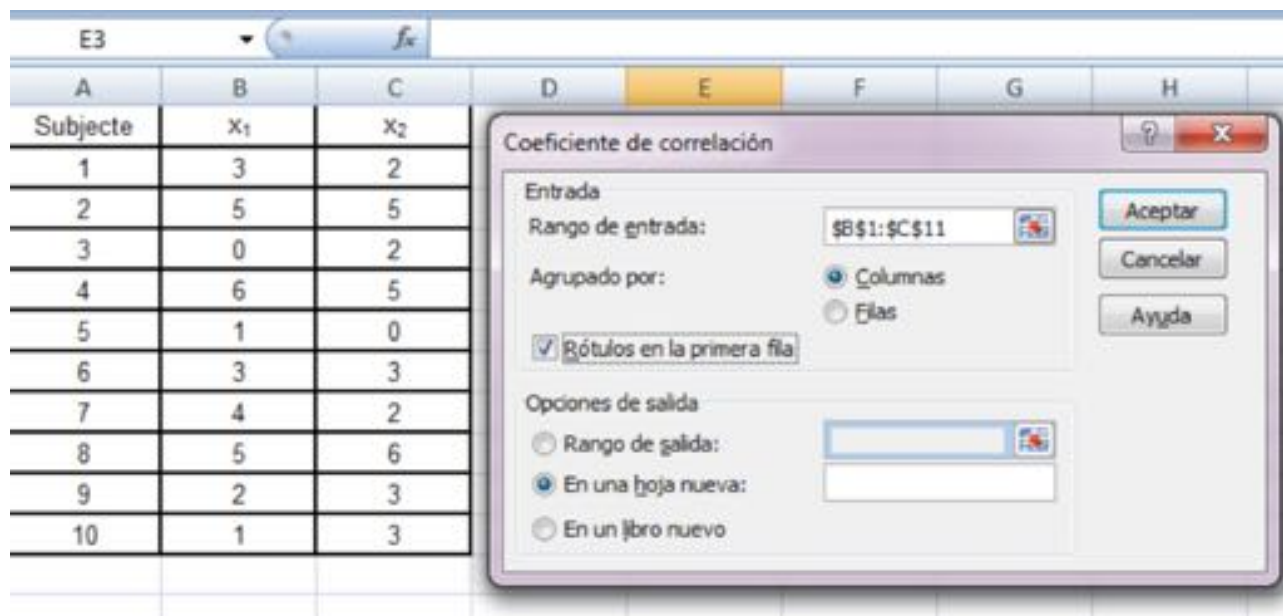


2) En el menú del **Coeficiente de correlación**, introduïu en el **Rango de entrada** el rang de les caselles on tingueu les dades de totes les variables, incloent-hi la primera filera amb les etiquetes.

3) Activeu l'opció **Rótulos en la primera fila**.

4) En **Opciones de salida**, mantingueu l'opció per defecte **En una hoja nueva**.

5) Premeu **Aceptar**.



6) En un nou full, us apareixerà el resultat del coeficient de correlació.

	A	B	C
1		x1	x2
2	x1	1	
3	x2	0,74406533	1
4			

2. Calculeu els indicadors d'acord següents entre les classificacions de les dues formes paral·leles:

- a) Coeficient de Hambleton i Novick (pH-N).
- b) Coeficient kappa (k).
- c) Coeficient de Livingston.

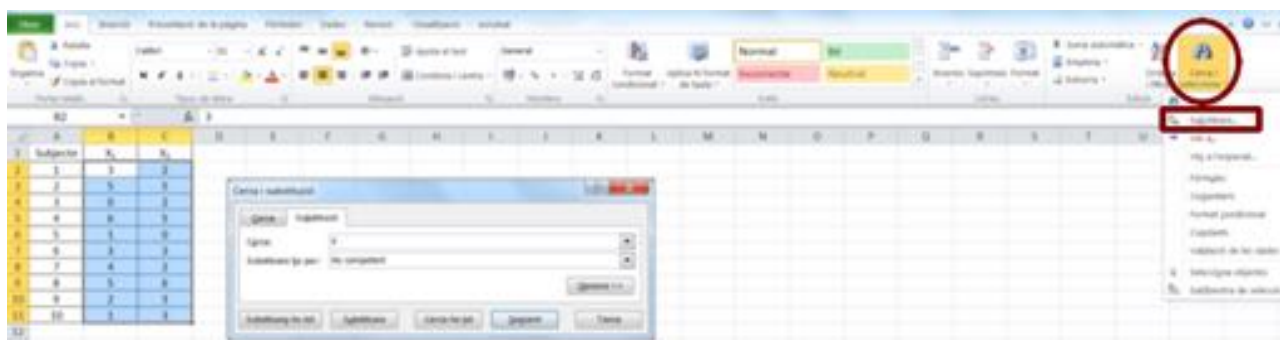
Per calcular els coeficients d'acord entre la classificació dels subjectes en les dues formes paral·leles d'aquest test, heu de generar una taula sabent que un subjecte és competent si té una puntuació superior a 2.

Per obtenir la taula amb l'Excel:

1) Activeu el rang de caselles on teniu els valors de les dues variables (B2 a C11).

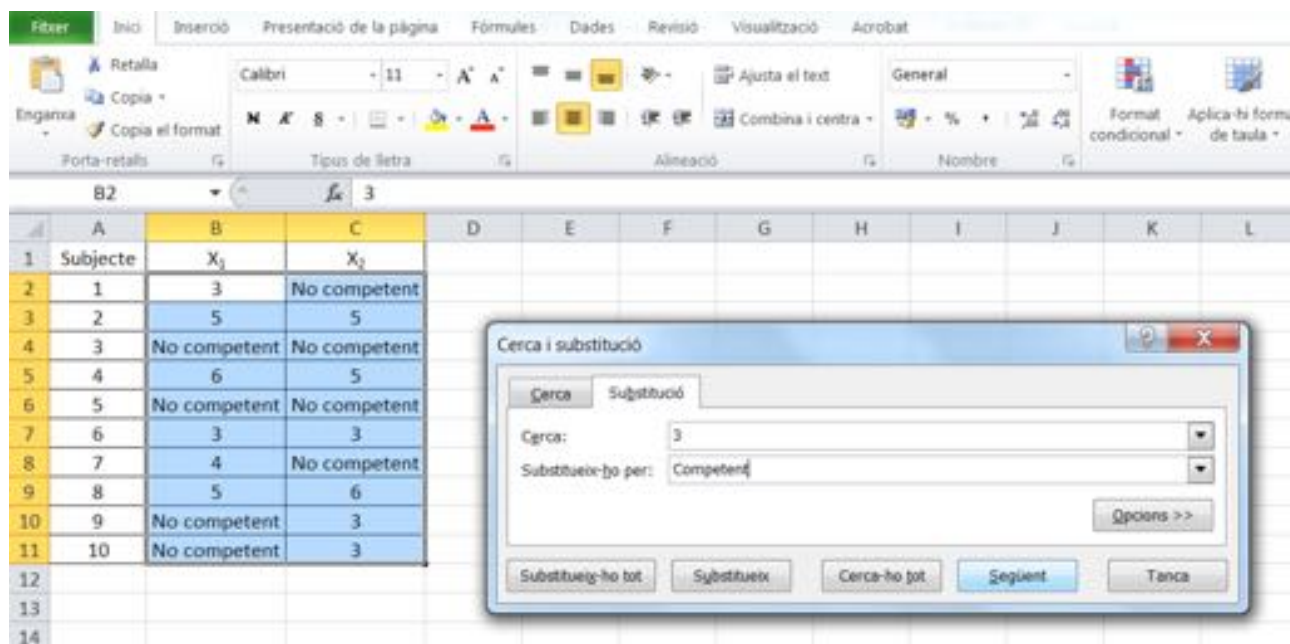
2) Seguiu la seqüència: **Cerca i selecciona, Substituir**.

3) En el quadre de diàleg introduïu un 0 a **Cercar**, i «No competent» a **Substitueix-ho per**. Cliqueu **Substitueix-ho tot**.



4) Repetiu el pas anterior, primer posant un 1 a **Cercar** i després posant un 2.

5) Repetiu els dos passos anteriors, però posant «Competent» a **Substitueix-ho per**, i 3, 4, 5 i 6 seguidament a l'espai de **Cercar**.

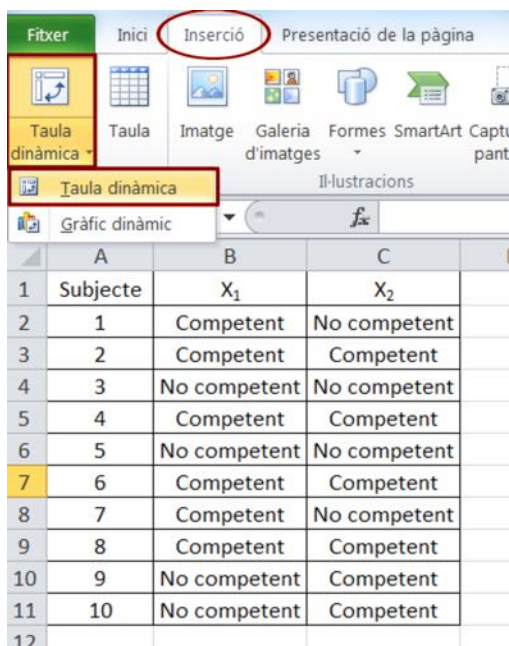


La taula resultant serà la següent:

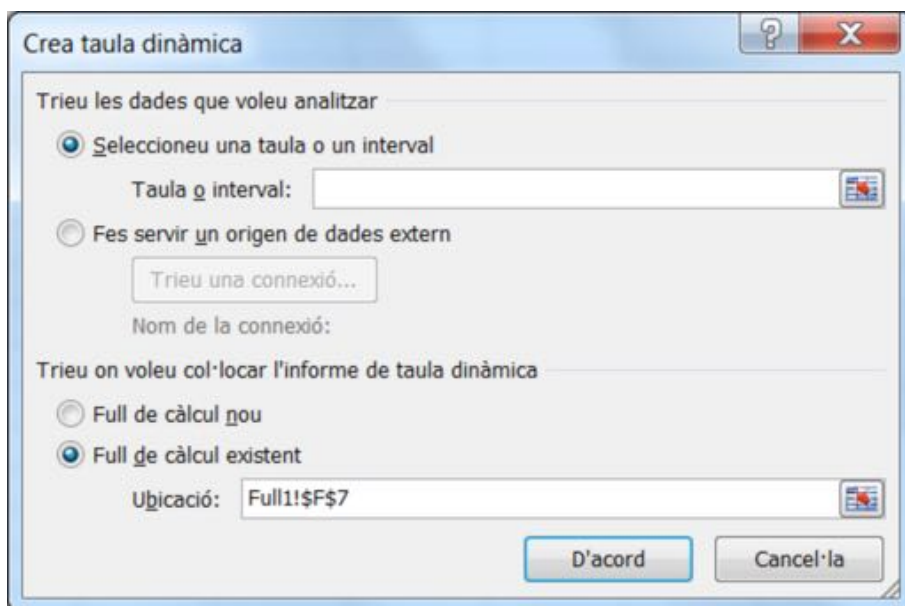
	A	B	C
1	Subjecte	X_1	X_2
2	1	Competent	No competent
3	2	Competent	Competent
4	3	No competent	No competent
5	4	Competent	Competent
6	5	No competent	No competent
7	6	Competent	Competent
8	7	Competent	No competent
9	8	Competent	Competent
10	9	No competent	Competent
11	10	No competent	Competent

A partir d'aquestes dades, heu de construir la taula de contingència, mitjançant l'opció de **Taula dinàmica**.

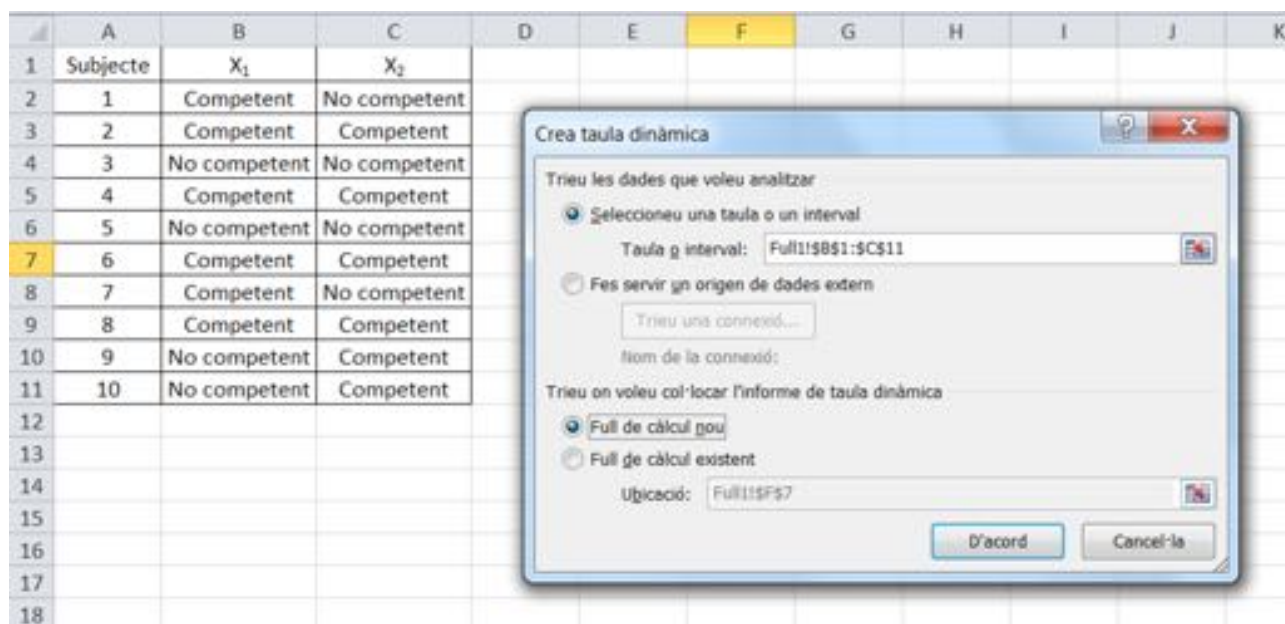
Per obtenir-la seguiu la seqüència: **Inserció, Taula dinàmica, Taula dinàmica**.



Us apareix el següent quadre de diàleg:



- 1) **Seleccioneu una taula o un interval:** seleccioneu les caselles on tingueu les dades anteriors.
- 2) Per col·locar la taula, marqueu **Full de càlcul nou**, així la tindreu en un nou full de l'Excel.
- 3) Cliqueu **D'acord**.



S'obre aquest nou full de l'Excel:

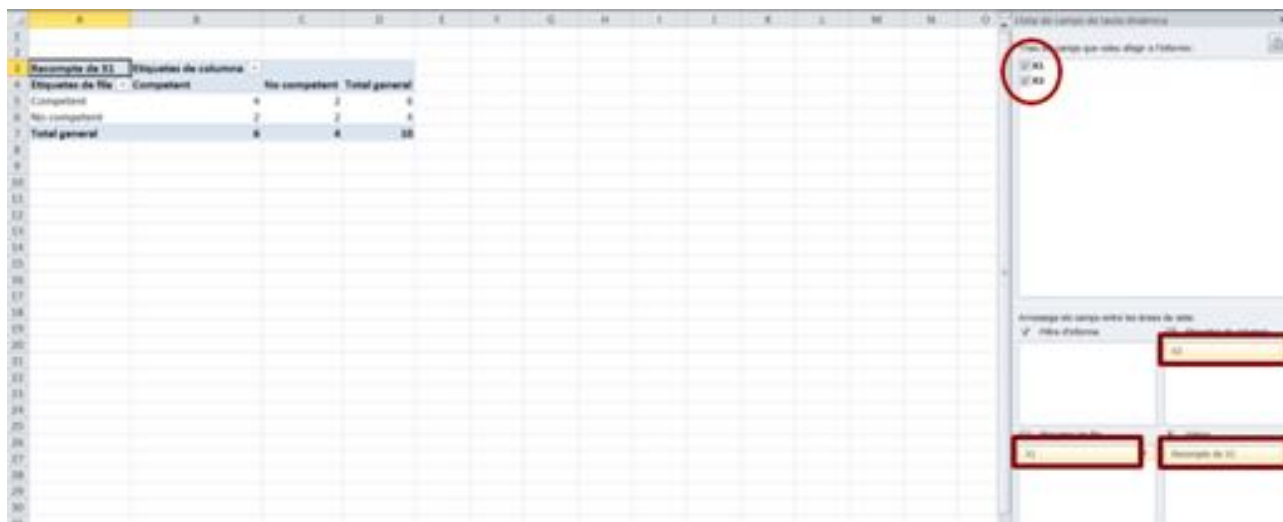


4) A la part esquerra teniu la taula per construir.

5) A la part dreta teniu els camps per omplir:

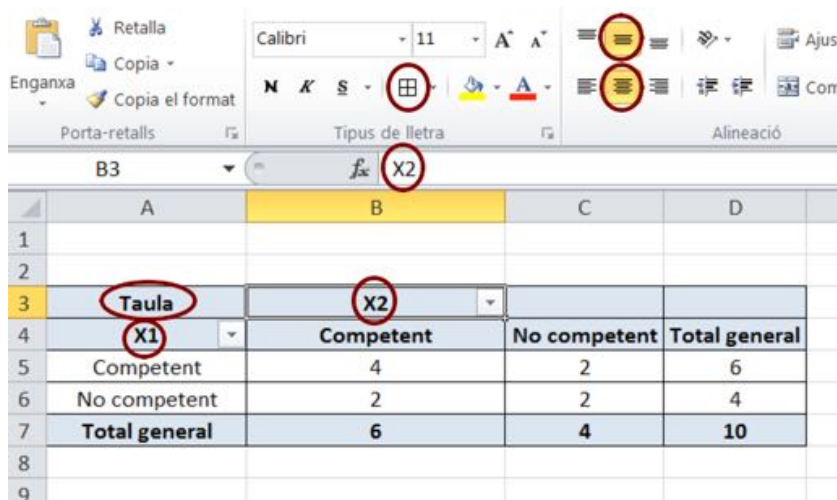
- Cliqueu sobre la x1 i arrossegueu el cursor fins a l'espai **Etiquetes de fila**.
- Cliqueu sobre la x2 i arrossegueu el cursor fins a l'espai **Etiquetes de columna**.
- Cliqueu de nou sobre la x1 i arrossegueu fins a l'espai **Σ Valors**.

La taula de contingència ens queda configurada a l'esquerra del full:



Podeu donar format a la taula amb les accions següents:

- Posar vores a les caselles.
- Centrar les caselles.
- Posar nom a la taula: «Taula».
- Posar etiquetes a les fileres (x1).
- Posar etiquetes a les columnes (x2).



Coeficient de Hambleton i Novick (p_{H-N})

La fórmula del coeficient de Hambleton i Novich és:

$$p_{H-N} = p_c - p_a$$

on, p_c és la proporció de classificacions consistents (6/10), i p_a la proporció de classificacions consistents que s'esperen per atzar i que s'obté amb la fórmula:

$$P_a = \sum \frac{n_{j \cdot} \cdot n_{i \cdot}}{N^2}$$

on:

- $n_{j \cdot}$ és el nombre de subjectes classificats com a competents (o no competents) per la forma x_1 ,
- $n_{i \cdot}$ és el nombre de subjectes classificats com a competents (o no competents) per la forma x_2 , i
- N és el nombre total de subjectes.

Per obtenir P_a amb l'Excel, seguiu els passos següents:

1) Marqueu la taula de contingència anterior i cliqueu **Copia**.

2) Us desplaceu a una casella buida (per exemple l'A12) i cliqueu amb el botó dret del ratolí **Enganxa-ho amb opcions** i en la secció **Enganxa els valors** seleccioneu l'opció **Format d'origen i valors**.

The screenshot shows the Excel interface with a context menu open over a cell. The menu includes options like 'Retalla', 'Copia', and 'Enganxa-ho amb opcions...'. The 'Enganxa-ho amb opcions...' option is selected, opening a sub-menu. In this sub-menu, the 'Enganxa els valors' section is expanded, and the 'Format d'origen i valors' option is highlighted. Below the menus, a contingency table is visible, which is the data being copied.

Taula	X2		
X1	Competent	No competent	Total general
Competent	4	2	6
No competent	2	2	4
Total general	6	4	10

Podeu donar format a la nova taula emmarcant les caselles i centrant-les.

Un cop teniu la nova taula, us col·loqueu en una casella buida (per exemple, B19) i introduïu la fórmula següent: $=((D14*B16)/(D16*D16))+((D15*C16)/(D16*D16))$, que us dona el resultat de 0,52.

B19		$f_x = ((D14*B16)/(D16*D16))+((D15*C16)/(D16*D16))$			
	A	B	C	D	E
1					
2					
3	Taula	X2			
4	X1	Competent	No competent	Total general	
5	Competent	4	2	6	
6	No competent	2	2	4	
7	Total general	6	4	10	
8					
9					
10					
11					
12	Taula	X2			
13	X1	Competent	No competent	Total general	
14	Competent	4	2	6	
15	No competent	2	2	4	
16	Total general	6	4	10	
17					
18					
19		0,52			
20					

I per obtenir el coeficient de Hambleton i Novich, us torneu a situar en una casella buida (per exemple, B21), i introduïu la fórmula següent: $=((B14+C15)/D16)-B19$, que us dona el resultat de 0,08.

B21		$f_x = ((B14+C15)/D16)-B19$			
	A	B	C	D	
1					
2					
3	Taula	X2			
4	X1	Competent	No competent	Total general	
5	Competent	4	2	6	
6	No competent	2	2	4	
7	Total general	6	4	10	
8					
9					
10					
11					
12	Taula	X2			
13	X1	Competent	No competent	Total general	
14	Competent	4	2	6	
15	No competent	2	2	4	
16	Total general	6	4	10	
17					
18					
19		0,52			
20					
21		0,08			
22					

$$p_{H-N} = p_c - p_a = 0,60 - 0,52 = 0,08$$

Pel coeficient kappa (K) i el coeficient de Livingston, simplement s'apliquen les fórmules corresponents amb la calculadora o amb l'Excel.

2.2. Exercici 2

Tenint en compte la informació i les dades del cas 2, contesteu els apartats següents en referència al test de depressió de sis ítems:

1. Obteniu la consistència interna (fiabilitat) del test utilitzant el mètode de les dues meitats a partir del següent:

- La fórmula d'Spearman Brown.
- La fórmula de Rulon.
- La fórmula de Gutman-Flanagan.

Fórmula d'Spearman Brown

Per obtenir la consistència interna o fiabilitat del test utilitzant el mètode de les dues meitats, a partir de la fórmula d'Spearman-Brown, seguiu els passos següents a partir de la matriu de dades inicials (CAS2, només amb les dades dels ítems 1 a 6):

1) Calculeu, per cada subjecte, el sumatori dels ítems imparells:

a) Us col·loqueu en la primera casella buida de la filera del primer subjecte (H2).

b) Introduïu la fórmula següent: =B2+D2+F2.

c) Premeu **Intro** i obtindreu el sumatori dels ítems imparells del primer subjecte en la casella H2.

H2		fx =B2+D2+F2						
	A	B	C	D	E	F	G	H
1	Subjecte	ítem 1	ítem 2	ítem 3	ítem 4	ítem 5	ítem 6	
2	1	1	1	0	0	0	1	
3	2	1	1	1	0	1	1	
4	3	0	1	1	0	0	0	
5								

d) Utilitzeu l'opció d'autoemplenament per emplenar la resta de caselles. Tal com s'ha explicat en l'apartat d'introducció als aspectes bàsics del programa Excel, col·loqueu-vos amb el cursor a l'angle inferior dret d'aquesta casella H2

fins que pareixi una creu (veieu l'imatge anterior), i cliqueu dos cops amb el ratolí (també podeu arrossegat la creu). D'aquesta manera, us apareixeran tots els resultats del sumatori dels ítems imparells dels vint-i-cinc subjectes.

A	B	C	D	E	F	G	H
Subjecte	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6	
1	1	1	0	0	0	1	1
2	1	1	1	0	1	1	3
3	0	1	1	0	0	0	1
4	0	1	1	0	0	0	1
5	0	0	1	1	1	1	2
6	0	0	0	1	1	1	1
7	0	0	1	0	1	1	2
8	0	0	1	0	0	0	1
9	1	1	1	0	1	1	3
10	0	0	1	0	1	1	2
11	0	1	1	0	0	1	1
12	1	1	0	1	1	1	2
13	1	0	1	1	1	1	3
14	0	0	0	0	0	0	0
15	1	0	0	0	1	1	2
16	1	0	1	0	0	0	2
17	1	1	1	1	1	1	3
18	0	1	0	0	1	1	1
19	1	1	1	1	1	1	3
20	0	1	1	0	1	1	2
21	0	1	1	0	1	1	2
22	0	0	0	0	1	0	1
23	0	1	1	0	1	1	2
24	0	1	1	0	0	1	1
25	0	0	0	0	1	0	1

2) Calculeu, per cada subjecte, el sumatori dels ítems parells:

a) Us col·loqueu en la nova primera casella buida de la filera del primer subjecte (I2).

b) Introduïu la fórmula següent: =C2+E2+G2.

c) Premeu **Intro** i obtindreu el sumatori dels ítems parells del primer subjecte en la casella I2.

I2								
=C2+E2+G2								
	A	B	C	D	E	F	G	H
1	Subjecte	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6	
2	1	1	1	0	0	0	1	1
3	2	1	1	1	0	1	1	3
4	3	0	1	1	0	0	0	1

d) Utilitzeu l'opció d'autoemplenament per emplenar la resta de caselles utilitzant la mateixa fórmula. Us apareixeran tots els resultats del sumatori dels ítems parells dels deu subjectes.

Podeu completar la nova taula encerclant les noves caselles i introduint les etiquetes de les variables generades «Imparells» i «Parells».

	A	B	C	D	E	F	G	H	I
1	Subjecte	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6	Imparells	Parells
2	1	1	1	0	0	0	1	1	2
3	2	1	1	1	0	1	1	3	2
4	3	0	1	1	0	0	0	1	1
5	4	0	1	1	0	0	0	1	1
6	5	0	0	1	1	1	1	2	2
7	6	0	0	0	1	1	1	1	2
8	7	0	0	1	0	1	1	2	1
9	8	0	0	1	0	0	0	1	0
10	9	1	1	1	0	1	1	3	2
11	10	0	0	1	0	1	1	2	1
12	11	0	1	1	0	0	1	1	2
13	12	1	1	0	1	1	1	2	3
14	13	1	0	1	1	1	1	3	2
15	14	0	0	0	0	0	0	0	0
16	15	1	0	0	0	1	1	2	1
17	16	1	0	1	0	0	0	2	0
18	17	1	1	1	1	1	1	3	3
19	18	0	1	0	0	1	1	1	2
20	19	1	1	1	1	1	1	3	3
21	20	0	1	1	0	1	1	2	2
22	21	0	1	1	0	1	1	2	2
23	22	0	0	0	0	1	0	1	0
24	23	0	1	1	0	1	1	2	2
25	24	0	1	1	0	0	1	1	2
26	25	0	0	0	0	1	0	1	0

e) Obtindrem el coeficient de correlació de Pearson entre el sumatori dels ítems parells i el dels imparells, amb la funció o amb el programa preconfigurat, tal com hem explicat en el primer exercici.

Amb el programa preconfigurat, obtenim el resultat següent en un nou full:

	A	B	C
1		Imparells	Parells
2	Imparells	1	
3	Parells	0,54656201	1
4			

3) Apliquem la fórmula d'Spearman Brown per obtenir el coeficient de fiabilitat del test:

$$r_{xx'} = \frac{2r_{pi}}{1+r_{pi}}$$

Amb l'Excel, us col·loqueu en una casella buida (per exemple la casella B6), i apliqueu la fórmula $= (2*B3)/(1+B3)$.

	A	B	C	D	E
1		Imparells	Parells		
2	Imparells	1			
3	Parells	0,546562	1		
4					
5					
6		0,706809			
7					

El coeficient de fiabilitat o consistència interna del test, aplicant la fórmula d'Spearman Brown és de 0,707.

Fórmula de Rulon

La fórmula de Rulon és:

$$r_{xx'} = 1 - \frac{S_d^2}{S_x^2}$$

on:

- S_d^2 : Variància de les diferències entre les puntuacions dels subjectes en les dues meitats del test.
- S_x^2 : Variància de les puntuacions totals dels subjectes en el test.

Per tant, necessitem les puntuacions totals dels subjectes en el test. Podeu calcular-les tal com heu fet amb els sumatoris dels ítems parells i els imparells.

	A	B	C	D	E	F	G	H	I	J
1	Subjecte	ítem 1	ítem 2	ítem 3	ítem 4	ítem 5	ítem 6	Imparells	Parells	
2	1	1	1	0	0	0	1	1	2	3
3	2	1	1	1	0	1	1	3	2	
4	3	0	1	1	0	0	0	1	1	
5	4	0	1	1	0	0	0	1	1	

I completar la taula amb el nom de la variable (Total).

	A	B	C	D	E	F	G	H	I	J
1	Subjecte	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6	Imparells	Parells	Total
2	1	1	1	0	0	0	1	1	2	3
3	2	1	1	1	0	1	1	3	2	5
4	3	0	1	1	0	0	0	1	1	2
5	4	0	1	1	0	0	0	1	1	2
6	5	0	0	1	1	1	1	2	2	4
7	6	0	0	0	1	1	1	1	2	3
8	7	0	0	1	0	1	1	2	1	3
9	8	0	0	1	0	0	0	1	0	1
10	9	1	1	1	0	1	1	3	2	5
11	10	0	0	1	0	1	1	2	1	3
12	11	0	1	1	0	0	1	1	2	3
13	12	1	1	0	1	1	1	2	3	5
14	13	1	0	1	1	1	1	3	2	5
15	14	0	0	0	0	0	0	0	0	0
16	15	1	0	0	0	1	1	2	1	3
17	16	1	0	1	0	0	0	2	0	2
18	17	1	1	1	1	1	1	3	3	6
19	18	0	1	0	0	1	1	1	2	3
20	19	1	1	1	1	1	1	3	3	6
21	20	0	1	1	0	1	1	2	2	4
22	21	0	1	1	0	1	1	2	2	4
23	22	0	0	0	0	1	0	1	0	1
24	23	0	1	1	0	1	1	2	2	4
25	24	0	1	1	0	0	1	1	2	3
26	25	0	0	0	0	1	0	1	0	1

Ara cal calcular les diferències entre els sumatoris dels ítems parells i els imparells obtinguts anteriorment. Aquest nova variable són les puntuacions de diferència (D).

K2											
	A	B	C	D	E	F	G	H	I	J	K
1	Subjecte	ítem 1	ítem 2	ítem 3	ítem 4	ítem 5	ítem 6	Imparells	Parells	Total	D
2	1	1	1	0	0	0	1	1	2	3	1
3	2	1	1	1	0	1	1	3	2	5	-1
4	3	0	1	1	0	0	0	1	1	2	0
5	4	0	1	1	0	0	0	1	1	2	0
6	5	0	0	1	1	1	1	2	2	4	0
7	6	0	0	0	1	1	1	1	2	3	1
8	7	0	0	1	0	1	1	2	1	3	-1
9	8	0	0	1	0	0	0	1	0	1	-1
10	9	1	1	1	0	1	1	3	2	5	-1
11	10	0	0	1	0	1	1	2	1	3	-1
12	11	0	1	1	0	0	1	1	2	3	1
13	12	1	1	0	1	1	1	2	3	5	1
14	13	1	0	1	1	1	1	3	2	5	-1
15	14	0	0	0	0	0	0	0	0	0	0
16	15	1	0	0	0	1	1	2	1	3	-1
17	16	1	0	1	0	0	0	2	0	2	-2
18	17	1	1	1	1	1	1	3	3	6	0
19	18	0	1	0	0	1	1	1	2	3	1
20	19	1	1	1	1	1	1	3	3	6	0
21	20	0	1	1	0	1	1	2	2	4	0
22	21	0	1	1	0	1	1	2	2	4	0
23	22	0	0	0	0	1	0	1	0	1	-1
24	23	0	1	1	0	1	1	2	2	4	0
25	24	0	1	1	0	0	1	1	2	3	1
26	25	0	0	0	0	1	0	1	0	1	-1

Per aplicar la fórmula de Rulon, cal obtenir les variàncies de les puntuacions de diferència (D), i de les puntuacions totals (Total).

Per obtenir aquestes variàncies amb l'Excel, cal utilitzar la funció **VARP**, seguint els següents passos:

- 1) Situeu el cursor en la primera casella buida de la columna de la variable Total (casella J27).
- 2) Seguiu la seqüència: **Fórmules**, **Insereix una funció**, seleccioneu la categoria **Estadístiques** i la funció **VAR.P**.

The screenshot shows the Microsoft Excel interface with the 'Formules' ribbon selected. The 'Insertar una funció' (Insert Function) dialog box is open, displaying the 'Estadístiques' (Statistics) category. The function 'VAR.P' is highlighted in the list. The background spreadsheet contains a table with the following data:

	A	B	C	D	E	F	G	H	I	J	K	L
1	Subjecte	ítem 1	ítem 2	ítem 3	ítem 4	ítem 5	ítem 6	Imparells	Parells	Total	D	
2	1	1	1	0	0	0	1	1	2	3	=I2-H2	
3	2	1	1	1	0	1	1	3	2	5	-1	
4	3	0	1	1	0	0	0	1	1	2	0	
5	4	0	1	1	0							
6	5	0	0	1	1							
7	6	0	0	0	1							
8	7	0	0	1	0							
9	8	0	0	1	0							
10	9	1	1	1	0							
11	10	0	0	1	0							
12	11	0	1	1	0							
13	12	1	1	0	1							
14	13	1	0	1	1							
15	14	0	0	0	0							
16	15	1	0	0	0							
17	16	1	0	1	0							
18	17	1	1	1	1							
19	18	0	1	0	0							
20	19	1	1	1	1							
21	20	0	1	1	0							
22	21	0	1	1	0							
23	22	0	0	0	0							
24	23	0	1	1	0							

3) Cliqueu **D'acord** i **D'acord** en el quadre de diàleg següent, i tindreu el valor de la variància de les puntuacions Totals en la casella J27 (2,4224).

4) Arrossegueu el cursor cap a la dreta a la columna de la variable D (amb el procediment d'autoemplenament), i us apareix la variància de les puntuacions de diferència (D) a la casella K27 (0,72).

	A	B	C	D	E	F	G	H	I	J	K
1	Subjecte	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6	Imparells	Parells	Total	D
2	1	1	1	0	0	0	1	1	2	3	1
3	2	1	1	1	0	1	1	3	2	5	-1
4	3	0	1	1	0	0	0	1	1	2	0
5	4	0	1	1	0	0	0	1	1	2	0
6	5	0	0	1	1	1	1	2	2	4	0
7	6	0	0	0	1	1	1	1	2	3	1
8	7	0	0	1	0	1	1	2	1	3	-1
9	8	0	0	1	0	0	0	1	0	1	-1
10	9	1	1	1	0	1	1	3	2	5	-1
11	10	0	0	1	0	1	1	2	1	3	-1
12	11	0	1	1	0	0	1	1	2	3	1
13	12	1	1	0	1	1	1	2	3	5	1
14	13	1	0	1	1	1	1	3	2	5	-1
15	14	0	0	0	0	0	0	0	0	0	0
16	15	1	0	0	0	1	1	2	1	3	-1
17	16	1	0	1	0	0	0	2	0	2	-2
18	17	1	1	1	1	1	1	3	3	6	0
19	18	0	1	0	0	1	1	1	2	3	1
20	19	1	1	1	1	1	1	3	3	6	0
21	20	0	1	1	0	1	1	2	2	4	0
22	21	0	1	1	0	1	1	2	2	4	0
23	22	0	0	0	0	1	0	1	0	1	-1
24	23	0	1	1	0	1	1	2	2	4	0
25	24	0	1	1	0	0	1	1	2	3	1
26	25	0	0	0	0	1	0	1	0	1	-1
27										2,4224	0,72

I apliqueu la fórmula de Rulon amb aquests valors:

$$r_{xx'} = 1 - \frac{S_d^2}{S_x^2} = 1 - \frac{0,72}{2,4224} = 0,703$$

El coeficient de fiabilitat o consistència interna del test, aplicant la fórmula de Rulon és de 0,703.

Fórmula de Gutman-Flanagan

La fórmula de Gutman-Flanagan calcula el coeficient de fiabilitat d'un test a partir de les variàncies dels ítems parells i imparells, i de la puntuació total. La seva expressió és:

$$r_{xx'} = 2 \left(1 - \frac{S_p^2 + S_i^2}{S_x^2} \right)$$

On:

- S_p^2 : Variància de les puntuacions dels subjectes en els ítems parells del test.
- S_i^2 : Variància de les puntuacions dels subjectes en els ítems imparells del test.

- S_x^2 : Variància de les puntuacions totals dels subjectes en el test.

Amb l'Excel calculeu aquestes variàncies amb la funció **VARP** comentada anteriorment.

	A	B	C	D	E	F	G	H	I	J	K
1	Subjecte	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6	Imparells	Parells	Total	D
2	1	1	1	0	0	0	1	1	2	3	1
3	2	1	1	1	0	1	1	3	2	5	-1
4	3	0	1	1	0	0	0	1	1	2	0
5	4	0	1	1	0	0	0	1	1	2	0
6	5	0	0	1	1	1	1	2	2	4	0
7	6	0	0	0	1	1	1	1	2	3	1
8	7	0	0	1	0	1	1	2	1	3	-1
9	8	0	0	1	0	0	0	1	0	1	-1
10	9	1	1	1	0	1	1	3	2	5	-1
11	10	0	0	1	0	1	1	2	1	3	-1
12	11	0	1	1	0	0	1	1	2	3	1
13	12	1	1	0	1	1	1	2	3	5	1
14	13	1	0	1	1	1	1	3	2	5	-1
15	14	0	0	0	0	0	0	0	0	0	0
16	15	1	0	0	0	1	1	2	1	3	-1
17	16	1	0	1	0	0	0	2	0	2	-2
18	17	1	1	1	1	1	1	3	3	6	0
19	18	0	1	0	0	1	1	1	2	3	1
20	19	1	1	1	1	1	1	3	3	6	0
21	20	0	1	1	0	1	1	2	2	4	0
22	21	0	1	1	0	1	1	2	2	4	0
23	22	0	0	0	0	1	0	1	0	1	-1
24	23	0	1	1	0	1	1	2	2	4	0
25	24	0	1	1	0	0	1	1	2	3	1
26	25	0	0	0	0	1	0	1	0	1	-1
27								0,6816	0,8896	2,4224	

Així, els valors de les variàncies són:

$$S_p^2 = 0,8890, S_i^2 = 0,6816, S_x^2 = 2,4224$$

i la fórmula de Gutman-Flanagan ens proporciona el resultat següent:

$$r_{xx'} = 2 \left(1 - \frac{S_p^2 + S_i^2}{S_x^2} \right) = 2 \left(1 - \frac{0,8896 + 0,6816}{2,4224} \right) = 0,703$$

El coeficient de fiabilitat o consistència interna del test, aplicant la fórmula de Gutman-Flanagan és de 0,703.

2. Obtingueu el coeficient alfa de Cronbach utilitzant les tres diferents fórmules per al seu càlcul, o sigui a partir de:

- La variància dels diferents ítems.
- La covariància entre els ítems.
- r_1 .

Coeficient alfa de Cronbach a partir de la variància dels diferents ítems

Per obtenir el coeficient α a partir de les variàncies dels diferents ítems, s'ha d'aplicar la fórmula següent:

$$\alpha = \frac{n}{n-1} \left[1 - \frac{\sum_{j=1}^n S_j^2}{S_x^2} \right]$$

on:

- n = Nombre d'ítems del test.
- $\sum_{j=1}^n S_j^2$ = Sumatori de les variàncies dels n ítems.
- S_x^2 = Variància de les puntuacions totals en el test.

Per tant, heu que calcular amb l'Excel les variàncies de cada ítem. La variància de les puntuacions totals ja l'heu calculat anteriorment.

	A	B	C	D	E	F	G	H	I	J	K
1	Subjecte	ítem 1	ítem 2	ítem 3	ítem 4	ítem 5	ítem 6	Imparells	Parells	Total	D
2	1	1	1	0	0	0	1	1	2	3	1
3	2	1	1	1	0	1	1	3	2	5	-1
4	3	0	1	1	0	0	0	1	1	2	0
5	4	0	1	1	0	0	0	1	1	2	0
6	5	0	0	1	1	1	1	2	2	4	0
7	6	0	0	0	1	1	1	1	2	3	1
8	7	0	0	1	0	1	1	2	1	3	-1
9	8	0	0	1	0	0	0	1	0	1	-1
10	9	1	1	1	0	1	1	3	2	5	-1
11	10	0	0	1	0	1	1	2	1	3	-1
12	11	0	1	1	0	0	1	1	2	3	1
13	12	1	1	0	1	1	1	2	3	5	1
14	13	1	0	1	1	1	1	3	2	5	-1
15	14	0	0	0	0	0	0	0	0	0	0
16	15	1	0	0	0	1	1	2	1	3	-1
17	16	1	0	1	0	0	0	2	0	2	-2
18	17	1	1	1	1	1	1	3	3	6	0
19	18	0	1	0	0	1	1	1	2	3	1
20	19	1	1	1	1	1	1	3	3	6	0
21	20	0	1	1	0	1	1	2	2	4	0
22	21	0	1	1	0	1	1	2	2	4	0
23	22	0	0	0	0	1	0	1	0	1	-1
24	23	0	1	1	0	1	1	2	2	4	0
25	24	0	1	1	0	0	1	1	2	3	1
26	25	0	0	0	0	1	0	1	0	1	-1
27		0,2304	0,2464	0,2176	0,1824	0,2176	0,2016			2,4224	0,72
28											

i ho apliqueu a la fórmula del coeficient α :

$$\alpha = \frac{n}{n-1} \left[1 - \frac{\sum_{j=1}^n S_j^2}{S_x^2} \right] = \frac{6}{5} \left(1 - \frac{0,2304 + 0,2464 + 0,2176 + 0,1824 + 0,2176 + 0,2016}{2,4224} \right) = 0,558$$

El coeficient α del test és de 0,558.

Coeficient alfa de Cronbach a partir de la covariància entre els ítems

La fórmula per obtenir el coeficient alfa a partir de les covariàncies entre els ítems és:

$$\alpha = \frac{n}{n-1} \left[\frac{\sum \sum_{j \neq k}^n \text{cov}(j, k)}{S_x^2} \right]$$

on:

- n = Nombre d'ítems del test.
- $\sum \sum \text{cov}(j, k)$ = Sumatori de les covariàncies dels n ítems.
- S_x^2 = Variància de les puntuacions en el test.

Per tant, heu d'obtenir les covariàncies entre els ítems.

Les covariàncies es poden obtenir mitjançant la funció (COVAR), o bé amb el programa preconfigurat corresponent.

Amb la funció:

- 1) Inserir les dades del Cas 2 en un full (només les dades corresponents als sis ítems).
- 2) Col·loqueu el cursor en una casella buida (per exemple, la I3).
- 3) Seguiu la seqüència: **Fórmules, Insereix una funció**.
- 4) Seleccioneu dins la categoria d'**Estadístiques** la funció **COVARIANÇA.P**.
- 5) Introduïu a **Matriu1** el rang de les caselles on teniu les dades de l'ítem 1 sense l'etiqueta de la primera filera B2:B26, i a **Matriu2** el rang de les caselles on teniu les dades de l'ítem 2 sense l'etiqueta de la primera filera: C2:C26.
- 6) Premeu **D'acord**.

Us apareix el resultat de la covariància entre l'ítem 1 i el 2 en la casella I3 (0,0384).

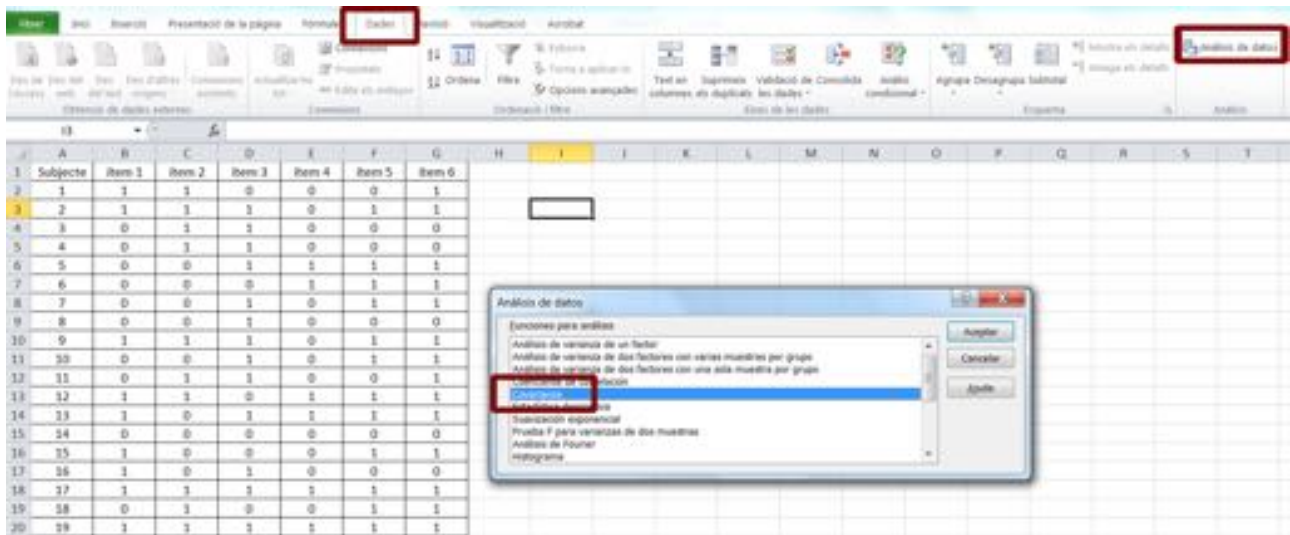
H	I	J
	0,0384	

Repetiu el procés per obtenir les covariàncies entre tota la resta de parelles d'ítems.

Amb el programa preconfigurat és molt més senzill obtenir la matriu de covariàncies entre tots els ítems:

1) Un cop heu activat el complement d'**Herramientas para análisis**, seguiu la seqüència **Dades, Análisis de datos, Covarianza**.

2) Premeu **Aceptar**.

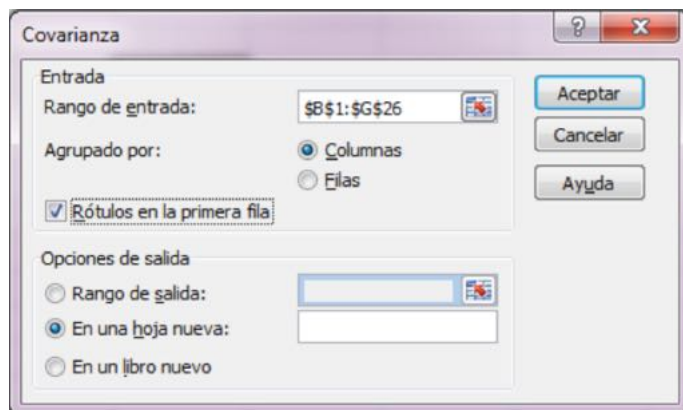


3) En el quadre de diàleg de **Covarianza**, introduïu en el **Rango de entrada** el rang de les caselles on tingueu les dades de totes les variables, incloent-hi la primera filera amb les etiquetes.

4) Activeu l'opció **Rótulos en la primera fila**.

5) En **Opciones de salida**, mantingueu l'opció per defecte **En una hoja nueva**.

6) Premeu **Aceptar**.



7) En un nou full, us apareixerà el resultat de la matriu de variàncies i covariàncies.

8) Les variàncies apareixen en la diagonal principal (encerclats en verd), i les covariàncies en la resta de caselles.

A	B	C	D	E	F	G
	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6
Ítem 1	0,2304					
Ítem 2	0,0384	0,2464				
Ítem 3	-0,0048	0,0592	0,2176			
Ítem 4	0,0736	-0,0144	-0,0032	0,1824		
Ítem 5	0,0352	-0,0208	-0,0224	0,0768	0,2176	
Ítem 6	0,0608	0,0768	0,0304	0,0672	0,1104	0,2016

Per obtenir el sumatori de les covariàncies, podeu situar-vos en una casella buida d'aquest full, i introduïu la següent fórmula: $=2*(B3+B4+B5+B6+B7+C7+C6+C5+C4+D5+D6+D7+E6+E7+F7)$, que és el sumatori de les covariàncies de la matriu multiplicat per dos, ja que aquesta matriu només mostra la meitat de les covariàncies. Així per exemple, només ens proporciona la covariància entre l'ítem 1 i el 2 (0,0384), però no la covariància entre l'ítem 2 i l'1, que és exactament la mateixa (0,0384).

	A	B	C	D	E	F	G	H	I
1		Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6		
2	Ítem 1	0,2304							
3	Ítem 2	0,0384	0,2464						
4	Ítem 3	-0,0048	0,0592	0,2176					
5	Ítem 4	0,0736	-0,0144	-0,0032	0,1824				
6	Ítem 5	0,0352	-0,0208	-0,0224	0,0768	0,2176			
7	Ítem 6	0,0608	0,0768	0,0304	0,0672	0,1104	0,2016		
8									
9									
10				1,1264					
11									

Per tant, el sumatori d'aquestes covariàncies serà igual a 1,1264, mentre que la variància de la puntuacions totals (S_x^2) ja sabem que és igual a 2,4224.

Per tant, a partir d'aquestes dades podeu calcular el coeficient α amb la calculadora o amb l'Excel a partir de la fórmula:

$$\alpha = \frac{n}{n-1} \left[\frac{\sum_{j \neq k}^n \text{cov}(j, k)}{S_x^2} \right] = \frac{6}{5} \left(\frac{1,1264}{2,4224} \right) = 0,558$$

El coeficient α del test és de 0,558, com ja havíem vist en l'apartat anterior.

Coefficient alfa de Cronbach a partir de r_1

Per obtenir el coeficient α de Cronbach a partir de r_1 només heu d'aplicar la fórmula corresponent amb les diferents dades que ja s'han calculat:

$$\alpha = \frac{n(r_1)}{1 + (n-1)r_1}$$

on:

- r_1 és el quocient entre la mitjana de les covariàncies i la mitjana de les variàncies dels diferents ítems del test, i
- n és el nombre d'ítems del test.

$$r_1 = \frac{1,1264/30}{1,296/6} = 0,1738$$

$$\alpha = \frac{n(r_1)}{1 + (n-1)r_1} = \frac{6 \times 0,1738}{1 + (5 \times 0,1738)} = 0,558$$

3. Determineu la significació estadística del coeficient alfa calculat en l'apartat anterior (nivell de confiança del 95%), i el seu interval de confiança.

Per tal de comprovar la significació estadística d'un coeficient alfa heu de seguir els passos següents:

1) Plantejament de les hipòtesis nul·la i l'alternativa:

- Hipòtesis nul·la: $\alpha = 0$
- Hipòtesis alternativa: $\alpha \neq 0$

2) Càlcul de l'estadístic de contrast amb la calculadora o amb l'Excel:

$$F = \frac{1 - \alpha}{1 - \hat{\alpha}} = \frac{1 - 0}{1 - 0,558} = 2,262$$

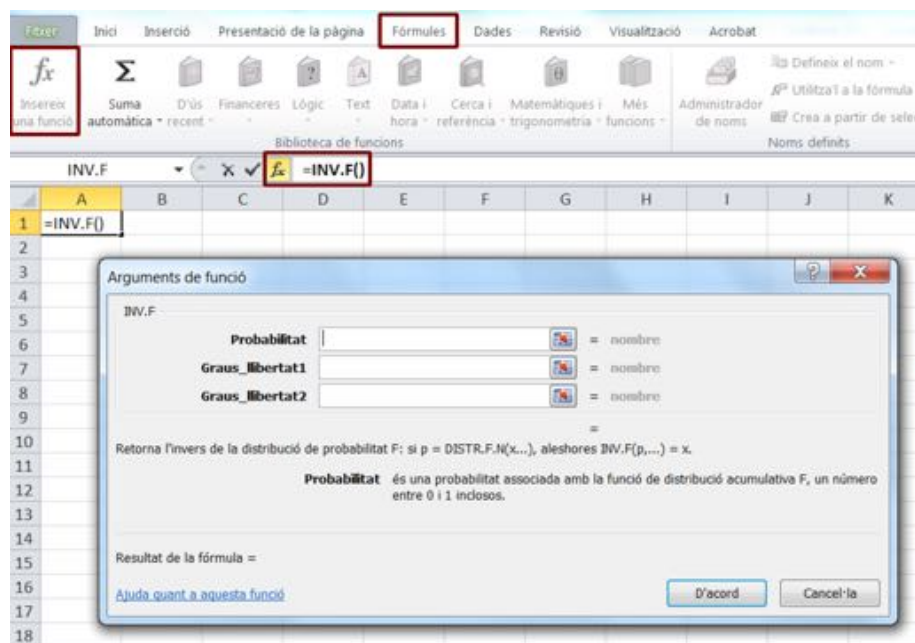
que es distribueix segons una F de Snedecor amb $(N-1)$ i $(N-1)(n-1)$ graus de llibertat, sent:

- N : Nombre de subjectes.
- n : Nombre d'ítems.
- α : Valor d'alfa en la població.
- $\hat{\alpha}$: Valor d'alfa calculat en la mostra.

3) Valors crítics de la distribució F de Snedecor amb 24 $(N-1)$ i 120 $((N-1)(n-1))$ graus de llibertat, per un nivell de confiança del 95% i contrast bilateral.

Per obtenir un valor crític de la distribució F de Snedecor amb l'Excel, heu d'utilitzar la funció **INV.F**, per tant heu de seguir la seqüència següent: **Fórmules, Insereix una funció** i dins la categoria **Estadístiques** cliqueu la funció **INV.F**.

S'obre el quadre de diàleg següent:



A **Probabilidad** heu d'introduir el valor de 0,975 per obtenir el valor crític inferior i 0,025 per obtenir el valor crític superior. Aquests valors són els adequats per un nivell de confiança del 95%. Si el nivell de confiança fos del 99%, els valors serien 0,995 i 0,005 respectivament.

A **Grados de libertad1** heu d'introduir els graus de llibertat $N-1$ (en el nostre cas, 24).

A **Grados de libertad2** heu d'introduir els graus de llibertat $(N-1)(n-1)$ (en el nostre cas, 120).

Argumens de funció

INV.F

Probabilitat 0,975 = 0,975

Graus_libertat1 24 = 24

Graus_libertat2 120 = 120

= 1,759725354

Retorna l'invers de la distribució de probabilitat F: si $p = \text{DISTR.F.N}(x, \dots)$, aleshores $\text{INV.F}(p, \dots) = x$.

Graus_libertat2 és el nombre de graus de llibertat del denominador, un número entre 1 i 10^{10} , exclos 10^{10} .

Resultat de la fórmula = 1,759725354

[Ajuda quant a aquesta funció](#) **D'acord** **Cancel·la**

Argumens de funció

INV.F

Probabilitat 0,025 = 0,025

Graus_libertat1 24 = 24

Graus_libertat2 120 = 120

= 0,49752579

Retorna l'invers de la distribució de probabilitat F: si $p = \text{DISTR.F.N}(x, \dots)$, aleshores $\text{INV.F}(p, \dots) = x$.

Probabilitat és una probabilitat associada amb la funció de distribució acumulativa F, un número entre 0 i 1 inclosos.

Resultat de la fórmula = 0,49752579

[Ajuda quant a aquesta funció](#) **D'acord** **Cancel·la**

Com podeu observar, els dos valors crítics són: 0,50 i 1,76 (arrodonint a dos decimals).

Com que el valor de l'estadístic de contrast obtingut (2,262) es troba fora del interval comprès entre els valors crítics 1,76 i 0,50, podem rebutjar la hipòtesi nul·la i podem concloure que, a partir de les nostres dades i amb un nivell de confiança del 95%, tenim evidència suficient per determinar que el valor del coeficient alfa en la població no és de zero, i per tant aquest coeficient és estadísticament significatiu.

Per construir l'interval confidencial d'aquest coeficient alfa, només cal substituir, en la fórmula de l'estadístic de contrast, els valors crítics de la distribució F i aïllar els valors d' α .

Aquests càlculs els podeu fer amb la calculadora o amb l'Excel.

$$\frac{1-\alpha}{1-0,558} \leq 1,76 \quad \alpha \geq 1-1,76(1-0,558)=0,22$$

$$\frac{1-\alpha}{1-0,558} \geq 0,50 \quad \alpha \leq 1-0,50(1-0,558)=0,78$$

Per tant, podem afirmar que amb un nivell de confiança del 95%, els valors del coeficient alfa en la població estaran compresos entre 0,22 i 0,78.

4. Si el mateix test s'administra a una mostra de trenta subjectes i s'obté un coeficient alfa de 0,70, hi ha diferències estadísticament significatives entre aquest coeficient alfa i l'obtingut en la mostra de vint-i-cinc subjectes del nostre exercici (nivell de confiança del 95%)?

Per comparar si la diferència entre dos coeficients alfa obtinguts en mostres diferents de subjectes són iguals o no, heu d'aplicar un contrast per dos coeficients en mostres independents. Per aplicar aquest contrast seguireu els passos següents:

1) Plantejament de les hipòtesis nul·la i alternativa:

- Hipòtesi nul·la: $\hat{\alpha}_1 = \hat{\alpha}_2$
- Hipòtesi alternativa: $\hat{\alpha}_1 \neq \hat{\alpha}_2$

2) Càlcul de l'estadístic de contrast amb la calculadora o amb l'Excel:

$$w = \frac{1-\hat{\alpha}_1}{1-\hat{\alpha}_2} = \frac{1-0,558}{1-0,70} = 1,47$$

que es distribueix segons una F de Snedecor amb (N_1-1) i (N_2-1) graus de llibertat, sent N_1 i N_2 les grandàries de les dues mostres.

3) Obtenció dels valors crítics de la distribució F de Snedecor amb 24 (N_1-1) i 29 (N_2-1) graus de llibertat, per un nivell de confiança del 95% i contrast bilateral amb la funció **DISTR.F.INV**, tal com hem vist en l'anterior apartat.

Arguments de funció

INV.F

Probabilitat	0,975	= 0,975
Graus_libertat1	24	= 24
Graus_libertat2	29	= 29

= 2,154005993

Retorna l'invers de la distribució de probabilitat F: si $p = \text{DISTR.F.N}(x, \dots)$, aleshores $\text{INV.F}(p, \dots) = x$.

Probabilitat és una probabilitat associada amb la funció de distribució acumulativa F, un número entre 0 i 1 inclosos.

Resultat de la fórmula = 2,154005993

[Ajuda quant a aquesta funció](#)

D'acord Cancel·la

Arguments de funció

INV.F

Probabilitat	0,025	= 0,025
Graus_libertat1	24	= 24
Graus_libertat2	29	= 29

= 0,45096993

Retorna l'invers de la distribució de probabilitat F: si $p = \text{DISTR.F.N}(x, \dots)$, aleshores $\text{INV.F}(p, \dots) = x$.

Probabilitat és una probabilitat associada amb la funció de distribució acumulativa F, un número entre 0 i 1 inclosos.

Resultat de la fórmula = 0,45096993

[Ajuda quant a aquesta funció](#)

D'acord Cancel·la

Com que el valor de l'estadístic de contrast obtingut (1,47) cau dins del interval comprès entre els valors crítics (0,45 – 2,15), no tenim prou evidències per rebutjar la hipòtesi nul·la, i per tant hem de concloure que la diferència entre els dos coeficients no és estadísticament significativa.

5. En una segona administració del test a la mateixa mostra inicial de vint-i-cinc subjectes, s'ha obtingut un coeficient alfa de 0,64 i una correlació entre les puntuacions dels subjectes en aquestes dues aplicacions de 0,92. Hi ha diferències estadísticament significatives entre els valors dels dos coeficients alfa (nivell de confiança del 95%)?

Contrast per dos coeficients en mostres dependents

Com hem administrat el mateix test en dues ocasions al mateix grup de subjectes per determinar si entre els dos coeficients alfa obtinguts hi ha o no diferències estadísticament significatives, aplicarem un contrast per dos coeficients en mostres dependents. Els passos que cal seguir són els següents:

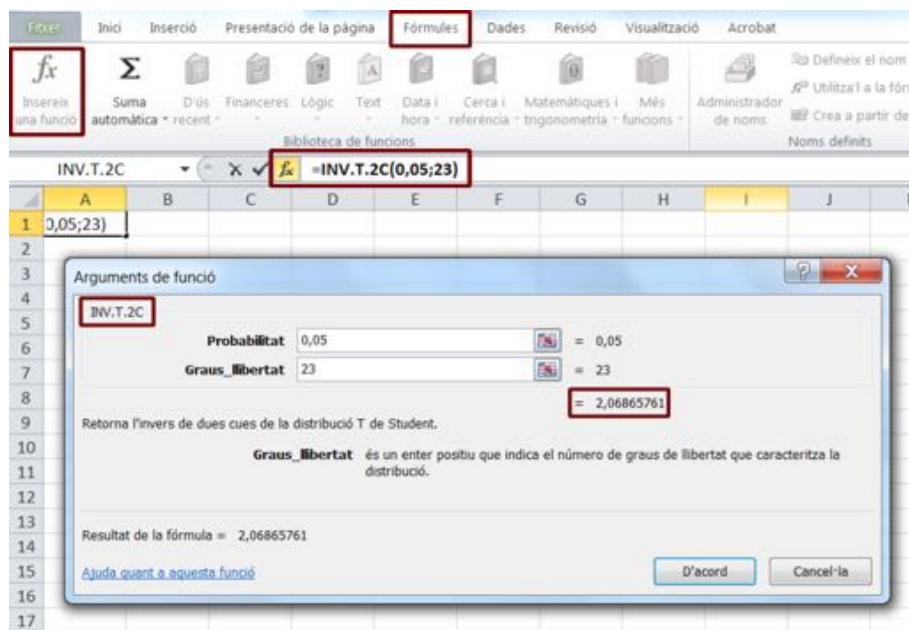
- 1) Hipòtesi nul·la: $\hat{\alpha}_1 = \hat{\alpha}_2$. Hipòtesi alternativa: $\hat{\alpha}_1 \neq \hat{\alpha}_2$.
- 2) Càlcul de l'estadístic de contrast amb la calculadora o amb l'Excel:

$$t = \frac{(\hat{\alpha}_1 - \hat{\alpha}_2)\sqrt{N-2}}{\sqrt{4(1-\hat{\alpha}_1)(1-\hat{\alpha}_2)(1-r_{12})}} = \frac{(0,558 - 0,64)\sqrt{25-2}}{\sqrt{4(1-0,558)(1-0,64)(1-0,92)}} = -1,74$$

que es distribueix segons una t de Student amb $N-2$ graus de llibertat, sent N el nombre de subjectes de la mostra.

Per obtenir amb l'Excel els valors crítics de la distribució t de Student amb 23 ($N-2$) graus de llibertat, per un nivell de confiança del 95% i contrast bilateral, s'utilitza la funció INV.T.2C.

Quan activeu aquesta funció, apareix el quadre de diàleg següent:



A **Probabilidad** heu d'introduir $1-0,95$ (1 menys el nivell de confiança en proporcions, o sigui 0,05 en el nostre cas), i a **Grados de libertad** els graus de llibertat de la t de Student ($N-2$ que en el nostre cas són 23).

El valor crític que ens proporciona aquesta funció és el superior, però com que la distribució t de Student és simètrica, el valor inferior és el mateix però en signe negatiu. Arrodonint a tres decimals, els dos valors són:

$$t_{0,975(23)} = 2,069 \text{ i } t_{0,025(23)} = -2,069$$

Com que l'estadístic de contrast obtingut (-1,74) queda dins del interval entre els valors crítics (-2,069 – 2,069), acceptem la hipòtesi nul·la, i podem concloure que amb un nivell de confiança del 95%, la diferència entre els dos coeficients alfa no és estadísticament significativa.

6. Calculeu el KR_{20} i el KR_{21} de Kuder-Richardson. Compareu els dos valors i raoneu el perquè de la igualtat o la diferència.

Per calcular KR_{20} podeu utilitzar el mateix procediment vist anteriorment pel coeficient α , ja que les variàncies de cada ítem no cal obtenir-les amb la fórmula dels productes p_j per q_j , sent (p_j) la proporció de subjectes que encerten l'ítem o tenen un 1 en aquest ítem i (q_j) la proporció de subjectes que no l'encerten o tenen un 0. Amb la funció **VAR.P** obtenim exactament el mateix resultat.

Per al càlcul de KR_{21} utilitzeu la fórmula corresponent, on només us caldrà obtenir la mitjana de les puntuacions totals en el test.

Per obtenir una mitjana amb l'Excel, utilitzeu la funció **MITJANA** de la categoria de funcions estadístiques.

En el cas de la mitjana de les puntuacions totals, aquesta mitjana és igual a 3,24.

	A	B	C	D	E	F
1	Total					
2	3					
3	5					
4	2					
5	2					
6	4					
7	3					
8	3					
9	1					
10	5					
11	3					
12	3					
13	5					
14	5					
15	0					
16	3					
17	2					
18	6					
19	3					
20	6					
21	4					
22	4					
23	1					
24	4					
25	3					
26	1					
27	3,24					
28						

7. Si afegíssim tres ítems més al test, suposant que mesuressin el mateix constructe, quina seria la seva nova fiabilitat?

8. Quants ítems hauríem d'afegir al test inicial per a arribar a una fiabilitat de 0,70?

Per als exercicis 7 i 8 només cal utilitzar les fórmules corresponents, i obtenir el resultat amb la calculadora o amb l'Excel.

9. Quina puntuació veritable podem estimar que tindrà el primer subjecte de la matriu de dades del nostre cas 1? Obteniu aquesta estimació amb un nivell de confiança del 95%, a partir del següent:

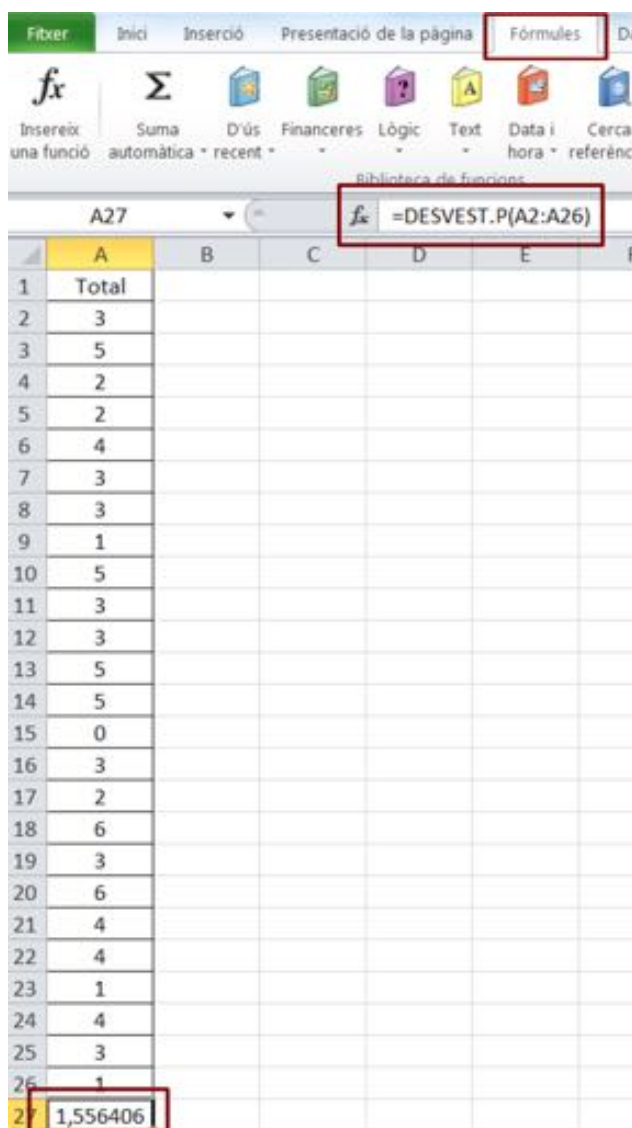
- La distribució normal dels errors.
- El model de la regressió.

En l'exercici 9, per fer l'estimació a partir de la distribució normal dels errors, només us falta saber com obtenir la desviació típica de les puntuacions totals, el valor de $Z_{\alpha/2}$ amb l'Excel, i l'arrel quadrada de $1 - 0,558$:

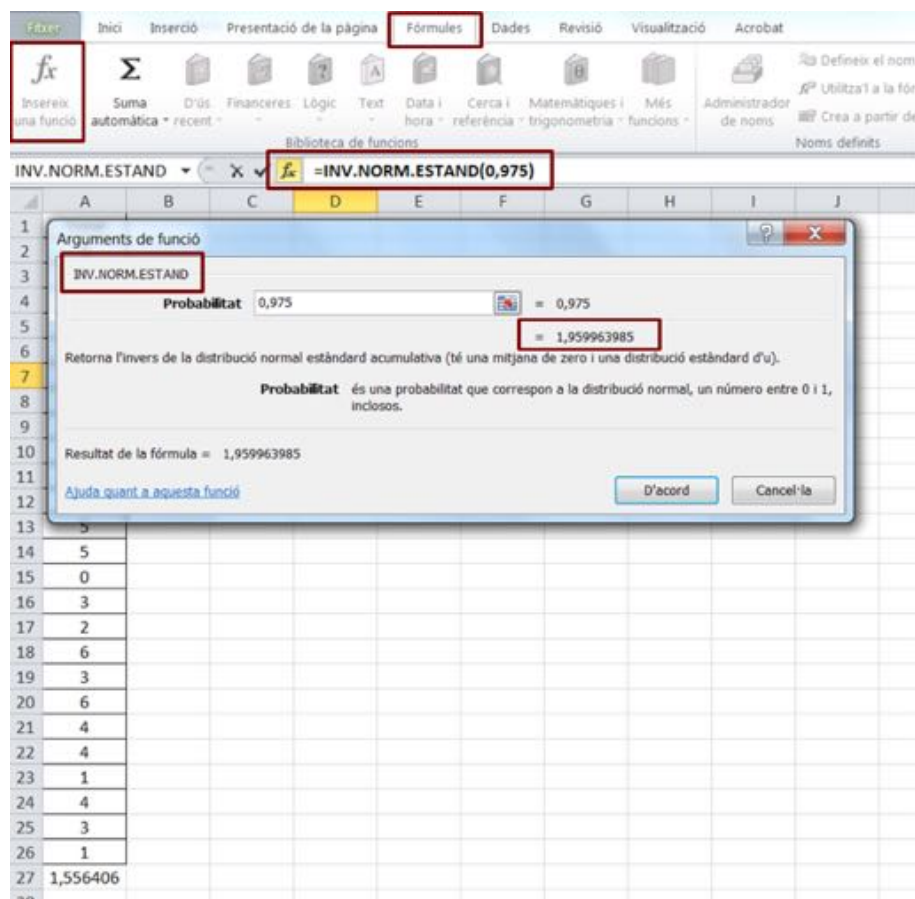
$$\left(\sqrt{1 - r_{xx}}\right)$$

Per obtenir una desviació típica, cal utilitzar la funció **DESVEST.P.**

En el nostre cas, la desviació típica de les puntuacions totals en el test serà d'1,5564.

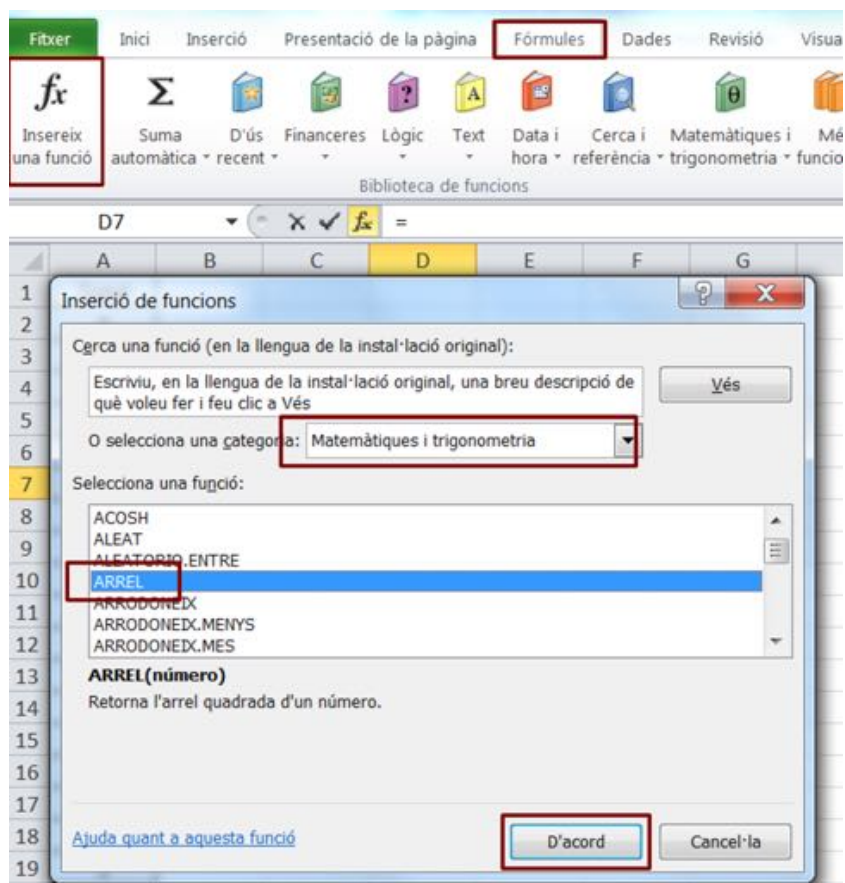


Per obtenir el valor $Z_{\alpha/2}$ per un nivell de confiança del 95%, utilitzeu la funció **INV.NORM.ESTAND**, introduint a **Probabilitat** el valor de 0,975 ($1 - \alpha/2$).



Com podem observar, el valor obtingut arrodonint a dos decimals és d'1,96.

Per calcular una arrel quadrada, cal utilitzar la funció **ARREL** que es troba dins la categoria de funcions **Matemàtiques i trigonometria**. Un cop activada aquesta funció, simplement cal introduir a **Número** el valor que volem extreure la seva arrel quadrada.



Per fer l'estimació a partir del model de la regressió utilitzant l'Excel, ja teniu tota la informació necessària en els apartats anteriors. Simplement és qüestió d'anar aplicant les fórmules i obtenir els resultats amb la calculadora o amb l'Excel.

3. Solucions dels exercicis de validesa

3.1. Exercici 1

A partir de les dades del cas 2, considerant les puntuacions del BDI com un criteri per a validar el nostre test de sis ítems, contesteu les qüestions següents:

1. Calculeu el coeficient de validesa del test.

En aquest cas per obtenir el coeficient de validesa del test, hem de calcular el coeficient de correlació de Pearson entre la variable puntuació total dels subjectes en el test i la puntuació en el criteri (BDI). Per obtenir un coeficient de correlació de Pearson amb l'Excel, podeu fer-ho amb la funció **COEF.CORREL** o amb el programa preconfigurat corresponent.

Amb la funció **COEF.CORREL**:

- 1) Inserir les dades del Cas 2 en un full.
- 2) Col·loqueu el cursor en una casella buida (per exemple, l'E3).
- 3) Seguiu la seqüència: **Fórmules, Insereix una funció**.
- 4) Seccioneu dins la categoria d'**Estadístiques** la funció **COEF.CORREL**.
- 5) Introduïu a **Matriu1** el rang de les caselles on teniu les dades de les puntuacions del test (TOTAL): B2:B26, i a **Matriu2** el rang de les caselles on teniu les dades del criteri (BDI): C2:C26.
- 6) Premeu **D'acord**.

The screenshot shows the Microsoft Excel interface with the 'Formules' ribbon selected. The formula bar displays the formula `=COEF.CORREL(B2:B26;C2:C26)`. A dialog box titled 'Argumentos de función' is open, showing the arguments for the COEF.CORREL function. Matrix1 is B2:B26 and Matrix2 is C2:C26. The result of the formula is 0,803528558.

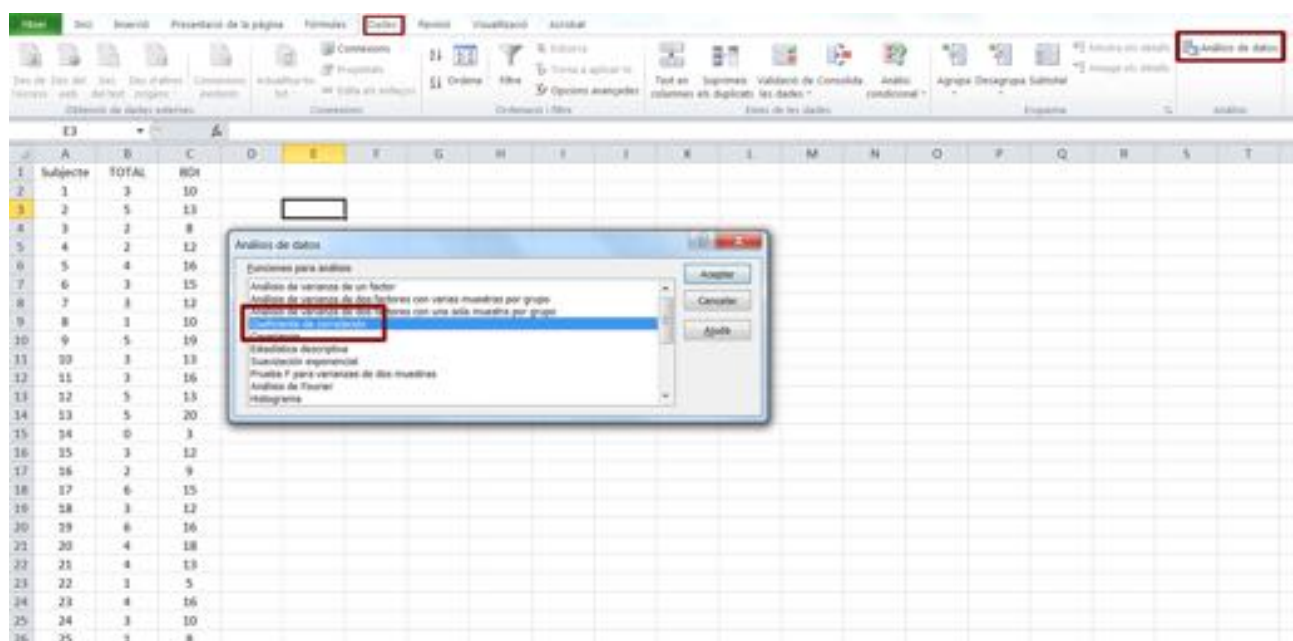
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Subjecte	TOTAL	BDI										
2	1	3	10										
3	2	5	13										
4	3	2	8										
5	4	2	12										
6	5	4	16										
7	6	3	15										
8	7	3	12										
9	8	1	10										
10	9	5	19										
11	10	3	13										
12	11	3	16										
13	12	5	13										
14	13	5	20										
15	14	0	3										
16	15	3	12										
17	16	2	9										
18	17	6	15										
19	18	3	12										
20	19	6	16										

A la casella E3 apareixerà el resultat del coeficient de correlació de Pearson que és de 0,803.

	A	B	C	D	E
1	Subjecte	TOTAL	BDI		
2	1	3	10		
3	2	5	13		0,803529
4	3	2	8		
5	4	2	12		
6	5	4	16		
7	6	3	15		
8	7	3	12		
9	8	1	10		
10	9	5	19		
11	10	3	13		
12	11	3	16		
13	12	5	13		
14	13	5	20		
15	14	0	3		
16	15	3	12		
17	16	2	9		
18	17	6	15		
19	18	3	12		
20	19	6	16		
21	20	4	18		
22	21	4	13		
23	22	1	5		
24	23	4	16		
25	24	3	10		
26	25	1	8		
27					

Amb el programa preconfigurat:

1) Un cop heu activat el complement d'**Herramientas para análisis**, seguiu la seqüència: **Dades**, **Análisis de datos**, **Coefficiente de correlación**.



2) En el menú del **Coeficiente de correlación**, introduïu en el **Rango de entrada** el rang de les caselles on tingueu les dades de totes les variables, incloent-hi la primera filera amb les etiquetes (\$B\$1:\$C\$26).

3) Activeu l'opció **Rótulos en la primera fila**.

4) En **Opciones de salida**, mantingueu l'opció per defecte **En una hoja nueva**.

5) Premeu **Aceptar**.

The screenshot shows the Excel interface with the 'Coeficiente de correlación' dialog box open. The spreadsheet data is as follows:

	A	B	C
1	Subjecte	TOTAL	BDI
2	1	3	10
3	2	5	13
4	3	2	8
5	4	2	12
6	5	4	16
7	6	3	15
8	7	3	12
9	8	1	10
10	9	5	19
11	10	3	13
12	11	3	16
13	12	5	13
14	13	5	20
15	14	0	3
16	15	3	12
17	16	2	9
18	17	6	15
19	18	3	12
20	19	6	16
21	20	4	18
22	21	4	13
23	22	1	5
24	23	4	16
25	24	3	10
26	25	1	8

The 'Coeficiente de correlación' dialog box settings are:

- Entrada**
 - Rango de entrada: \$B\$1:\$C\$26
- Agrupado por:**
 - ☒ Columnas
 - ☐ Filas
- ☒ Rótulos en la primera fila
- Opciones de salida**
 - ☐ Rango de salida:
 - ☒ En una hoja nueva:
 - ☐ En un libro nuevo

6) En un nou full, us apareixerà el resultat del coeficient de correlació.

	A	B	C
1		TOTAL	BDI
2	TOTAL	1	
3	BDI	0,80352856	1
4			

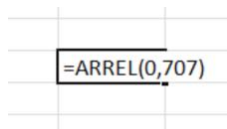
2. Quin seria aquest coeficient de validesa si el test tingués una fiabilitat perfecta (tingueu en compte la fiabilitat del test obtinguda a partir del mètode de les dues meitats i la fórmula d'Spearman Brown calculada en l'exercici 2 de l'apartat 2.2)?

En aquest cas caldrà aplicar la fórmula d'atenuació següent:

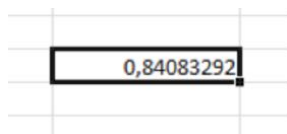
$$r_{v_{xy}} = \frac{r_{xy}}{\sqrt{r_{xx'}}}$$

En aquest cas podeu utilitzar l'Excel com si fos una calculadora:

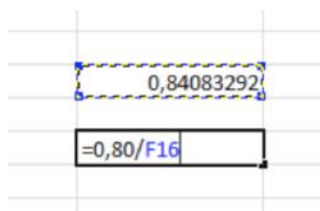
1) Activeu les funcions de l'Excel escrivint el signe d'igualtat (=) en qualsevol casella i indiqueu que voleu fer l'arrel quadrada del valor de la fiabilitat de la manera següent:



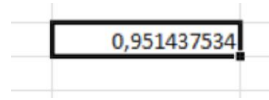
2) Premeu la tecla  (intro) i us pareixerà el resultat de l'arrel quadrada.



3) A continuació, calculeu el quocient entre el valor de la validesa obtingut al punt 1 i el resultat de l'arrel quadrada (podeu seleccionar la casella on apareix el resultat de l'arrel quadrada, en el nostre cas casella F16, per tal de no haver de tornar a escriure el resultat):



4) Si premeu  us apareixerà el resultat del coeficient de validesa del test suposant-hi una fiabilitat perfecta:



3. Quant valdria el coeficient de validesa del test si hi afegíssim sis ítems més? Utilitzeu el mateix coeficient de fiabilitat de l'apartat anterior.

Utilitzant l'expressió següent i de nou utilitzant l'Excel com una calculadora podreu obtenir la nova validesa en cas que afegíssim sis ítems més al test.

$$r_{xy} = \frac{r_{xy}\sqrt{n}}{\sqrt{1+(n-1)r_{xx'}}}$$

On:

- r_{xy} : és el valor inicial de la correlació test-criteri (en el nostre cas $r_{xy} = 0,80$)
- $r_{xx'}$: és el coeficient de fiabilitat del test (en el nostre cas $r_{xx'} = 0,707$)
- n : és el nombre de vegades que s'augmenta el test (en el nostre cas $n = 2$)

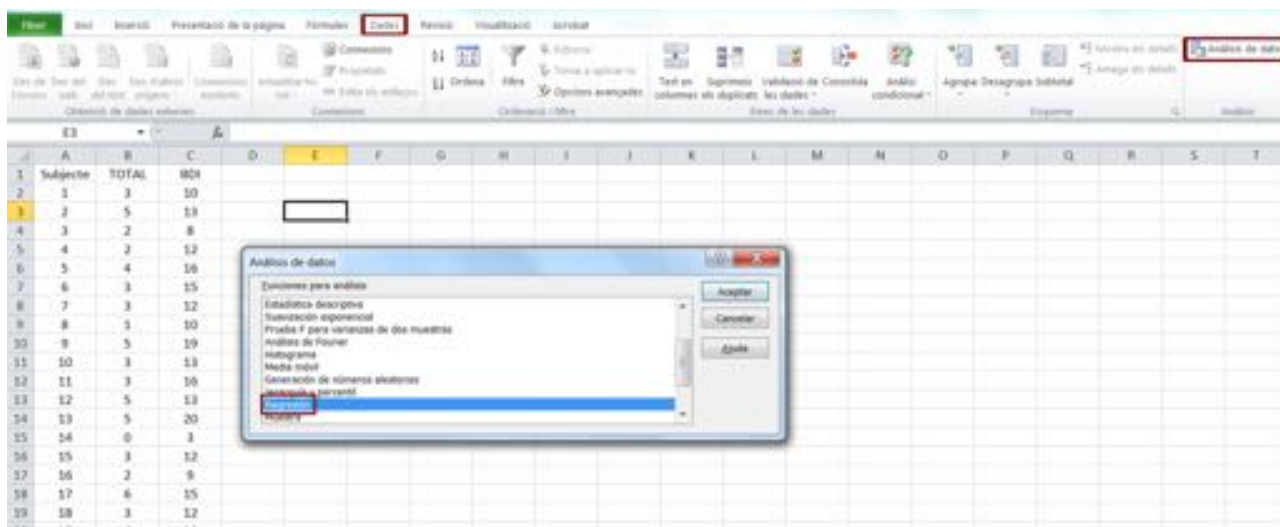
$$n = \frac{6+6}{2} = 2$$

$$r_{xy} = \frac{r_{xy}\sqrt{n}}{\sqrt{1+(n-1)r_{xx'}}} = \frac{0,80\sqrt{2}}{\sqrt{1+(2-1)0,707}} = 0,84$$

4. Entre quines puntuacions en el criteri (BDI) podem pronosticar que obtindrà la puntuació un nou subjecte que té 4 punts en el nostre test? Construïu l'interval amb un nivell de confiança del 95%.

Per calcular la puntuació que obtindria un subjecte en el criteri (BDI) sabent que ha obtingut una puntuació de 4 punts en el test ($x = 4$) aplicarem el model de regressió utilitzant el programa preconfigurat de l'Excel.

1) Un cop heu activat el complement d'**Herramientas para análisis**, seguiu la seqüència: **Dades, Análisis de datos, Regresión** i premeu **Aceptar**.



2) Un cop en el menú de **Regresión**, introduïu en el **Rango Y de entrada** el rang de les caselles on tingueu les dades de la variable criteri (BDI), incloent-hi la primera filera amb el nom de la variable.

3) Introduïu en el **Rango X de entrada** el rang de les caselles on tingueu les dades de la variable predictora (TOTAL), incloent-hi la primera filera amb el nom de la variable.

4) Activeu l'opció **Rótulos**.

5) Activeu l'opció **Nivel de confianza** i mantingueu per defecte el valor 95%.

6) En **Opciones de salida**, mantingueu l'opció per defecte **En una hoja nueva**.

7) Premeu **Aceptar**.

Regresión

Entrada

Rango Y de entrada:

Rango X de entrada:

☒ Rótulos ☐ Constante igual a cero

☒ Nivel de confianza %

Opciones de salida

☐ Rango de salida:

☒ En una hoja nueva:

☐ En un libro nuevo

Residuales

☐ Residuos ☐ Gráfico de residuales

☐ Residuos estándares ☐ Curva de regresión ajustada

Probabilidad normal

☐ Gráfico de probabilidad normal

Aceptar Cancelar Ayuda

El resultat que ens apareixerà en un full nou serà el següent:

Resumen							
Estadísticas de la regresión							
Coeficiente de cor	0,803528558						
Coeficiente de det	0,645658143						
R^2 ajustado	0,630251975						
Error típico	2,513760876						
Observaciones	25						

ANÁLISIS DE VARIANZA					
	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor crítico de F
Regresión	1	264,823144	264,823144	41,9090689	1,32013E-06
Residuos	23	145,336856	6,31899374		
Total	24	410,16			

	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%	Inferior 95,0%	Superior 95,0%
Intercepción	5,784676354	1,161080379	4,982149778	4,8754E-05	3,382798591	8,18655412	3,382798591	8,186554117
TOTAL	2,091149273	0,323021202	6,47372141	1,3201E-06	1,422929005	2,75936954	1,422929005	2,759369542

A la tercera taula, en l'apartat de coeficients, el valor corresponent a «Intercepción» representarà el valor de la intercepció (a), mentre que el valor del coeficient associat a TOTAL serà el pendent de la recta (b).

Així, a partir de la recta de regressió:

$$y' = a + bx$$

I sabent que el valor de x és $x = 4$, podrem calcular el valor pronosticat en la variable criteri (BDI):

$$y' = 5,78 + 2,09 \cdot 4$$

$$y' = 14,15$$

A partir d'aquest valor podrem calcular l'interval de confiança utilitzant l'expressió següent:

$$IC^{1-\alpha} \rightarrow y' \pm t_{n-1;\alpha/2} \cdot s_{y-y'}$$

On:

- $t_{n-1;\alpha/2}$ és el valor de la t de Student amb $n-1$ graus de llibertat (en el nostre cas $g.l.l = 24$) i el nivell de confiança al 95%, i
- $s_{y-y'}$ és l'error típic d'estimació.

8) A partir de la consulta de les taules de la t d'Student comproveu que $t = 2,064$. No obstant això, aquest valor també el podem obtenir a partir de l'Excel seleccionant la funció **INV.T.2C**. Aquesta funció ens proporcionarà el valor de t d'Student en funció del nivell de confiança (o bé del risc alfa amb el qual volem treballar) i graus de llibertat. Els passos per obtenir el valor de t de Student serien els següents:

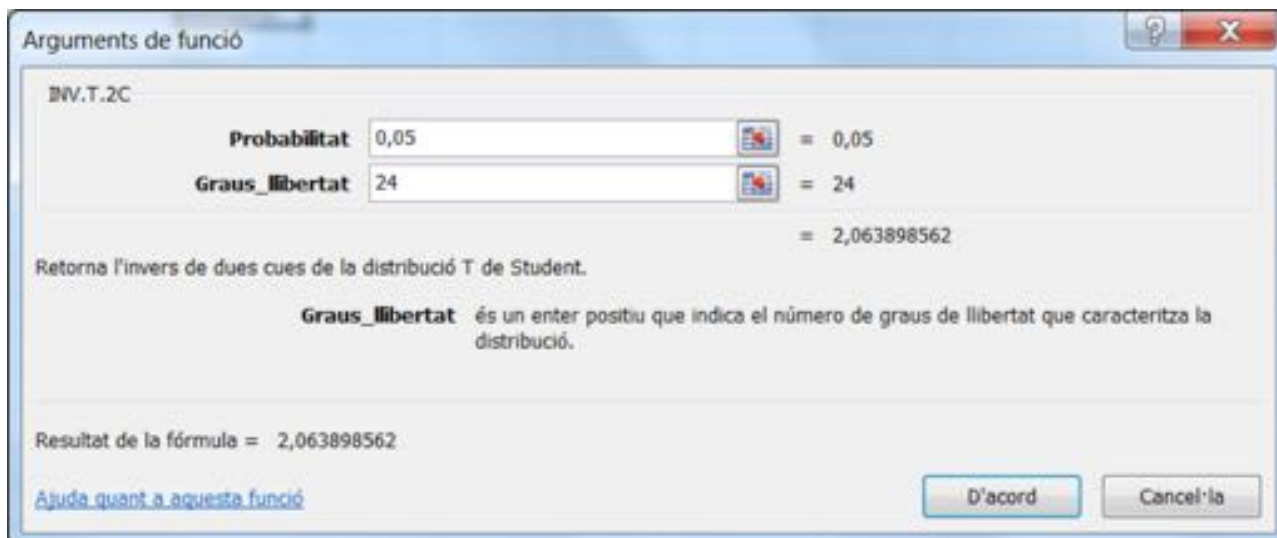
a) Activeu les funcions preconfigurades a partir de l'icona fx de la barra de fórmules: .

b) Seleccioneu la funció **INV.T.2C** dins de la categoria **Estadístiques**.

c) Introduïu la probabilitat associada a la distribució t de Student en el camp **Probabilitat**. En el nostre cas, com que volem treballar amb un nivell de confiança del 95%, introduïrem la probabilitat de 0,05.

d) Introduïu els graus de llibertat associats a la prova en el camp **Graus de llibertat**. En el nostre cas, introduïu el valor 24.

e) Premeu **D'acord** i el valor de la distribució t d'Student us apareixerà en una casella dins del full de càlcul.

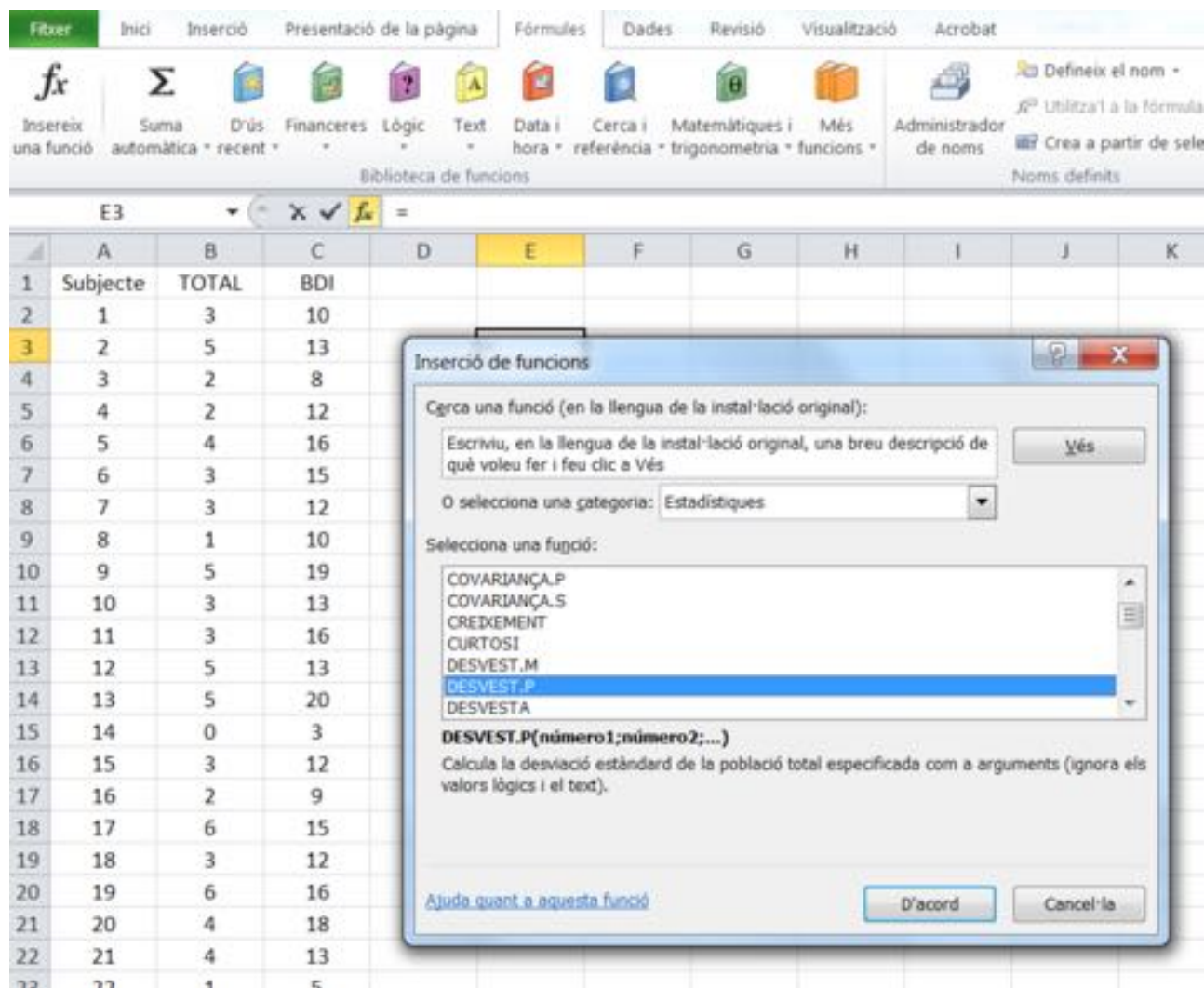


9) Calculeu l'error típic a partir de l'expressió següent:

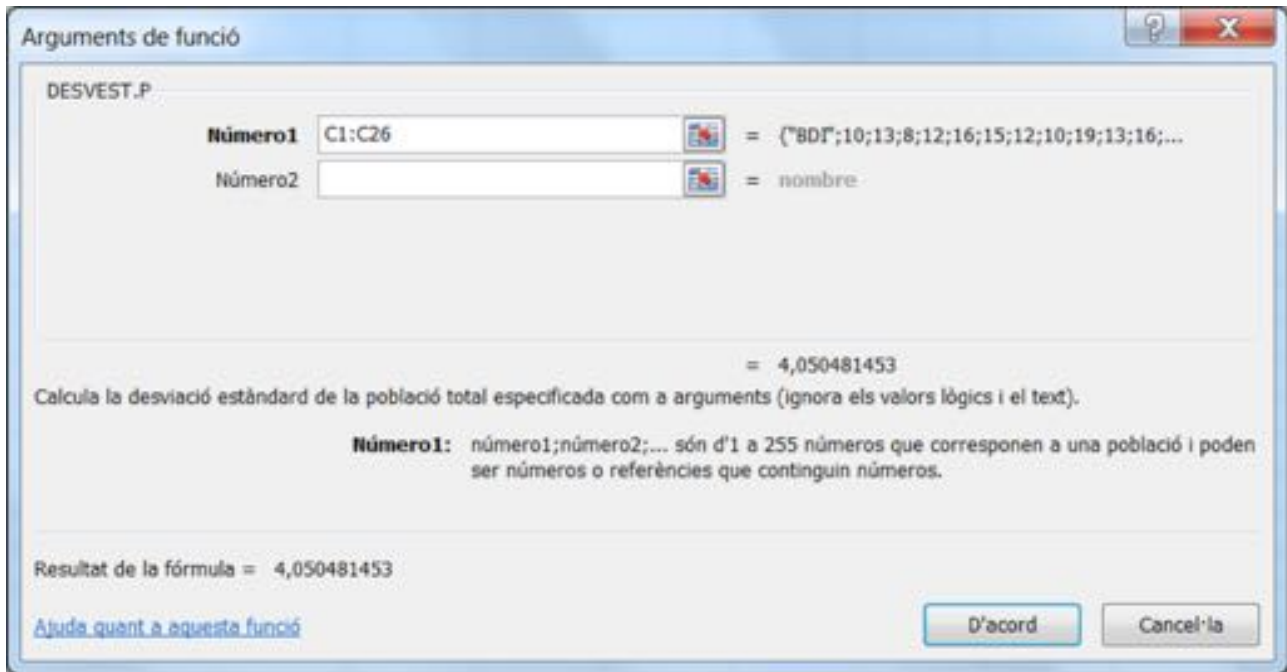
$$s_{y-y'} = s_y \sqrt{1 - r^2}$$

10) Per poder calcular l'error típic ens cal la desviació típica (S_y) del criteri, la qual obtindrem utilitzant la funció **DESVEST.P**. Per això, situeu el cursor en una casella buida (per exemple, la casella E3).

11) Seguiu la seqüència: **Fórmules**, **Insereix una funció**, seleccioneu la categoria **Estadístiques**, la funció **DESVEST.P** i premeu **D'acord**.



12) En la nova pantalla que us apareix, seleccioneu el rang d'entrada de la variable criteri (BDI) i premeu **D'acord**. Us apareixerà a la casella E3 el valor de la desviació típica de la variable BDI.



13) Una vegada heu obtingut S_y , podeu substituir la resta de valors de la fórmula i calcular l'error típic utilitzant l'Excel com una calculadora.

$$s_{y-y'} = 4,05\sqrt{1 - 0,80^2}$$

$$s_{y-y'} = 2,43$$

14) Finalment, podeu calcular l'interval de confiança utilitzant l'Excel com una calculadora.

$$IC^{1-\alpha} \rightarrow 10,15 \pm 2,064 \cdot 2,43 = [9,12; 19,15]$$

Podem concloure que, a un subjecte que ha obtingut 4 punts en el test, li podem pronosticar que tindrà entre 9 i 19 punts (arrodonint) en el BDI, amb un nivell de confiança del 95%.

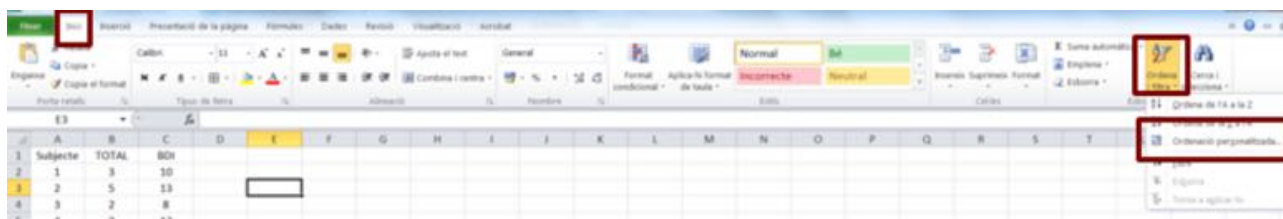
5. Si suposem que el diagnòstic ja contrastat de trastorn depressiu lleu se situa en una puntuació superior a 15 en el BDI, i nosaltres volem comprovar la validesa de la decisió del nostre test de sis ítems, suposant que puntuacions superiors a tres són les que detectarien aquest trastorn depressiu lleu, calculeu i interpreteu:

- El percentatge d'acord entre els dos tests.
- El coeficient kappa.
- La sensibilitat del nostre test.
- L'especificitat del nostre test.

Per poder calcular la validesa de decisió del nostre test haurem, en primer lloc, de recategoritzar la variable TOTAL i la variable BDI i convertir-les en variables dicotòmiques amb dues categories («no trastorn», «sí trastorn»).

En el cas de la variable TOTAL els valors iguals o inferiors a tres indicaran «no trastorn» i valors superiors a tres indicaran «sí trastorn». Per fer això copiarem els valors de la variable TOTAL en una columna nova, li canviarem el nom (per exemple, la podem anomenar TOTAL_R) i ordenarem de menor a major els valors de la variable. Per ordenar els valors seguirem els passos següents:

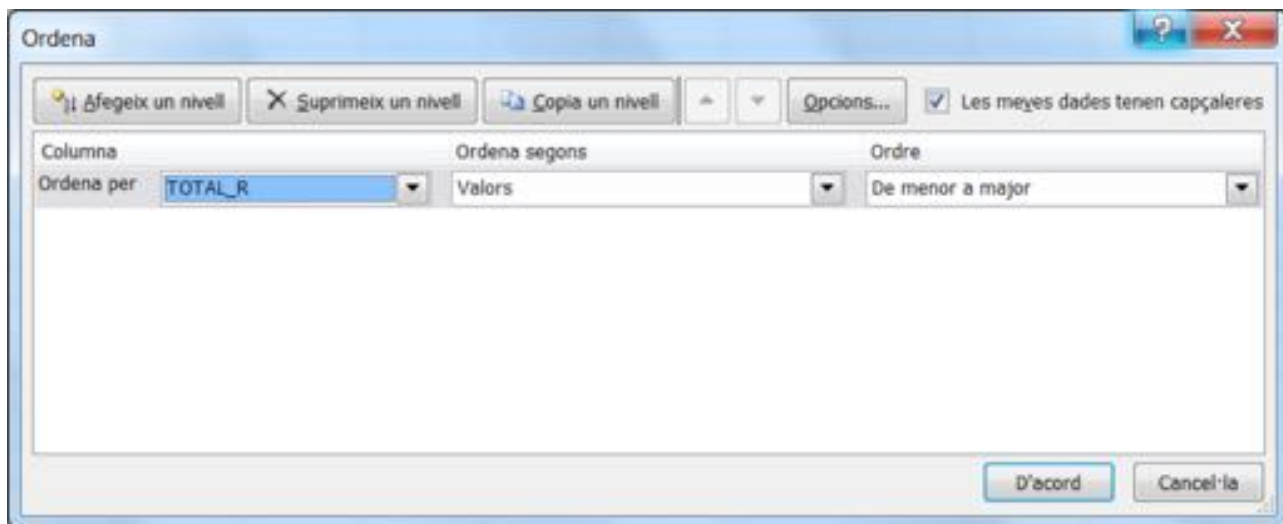
1) Seguiu la seqüència: **Inici, Ordena i filtra i Ordenació personalitzada.**



2) Seleccioneu dins de la categoria **Ordena per** la variable «TOTAL_R».

3) Seleccioneu com a criteri d'ordenació **De menor a major**.

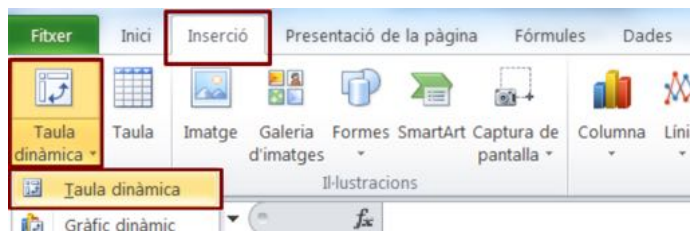
4) Premeu **D'acord**.



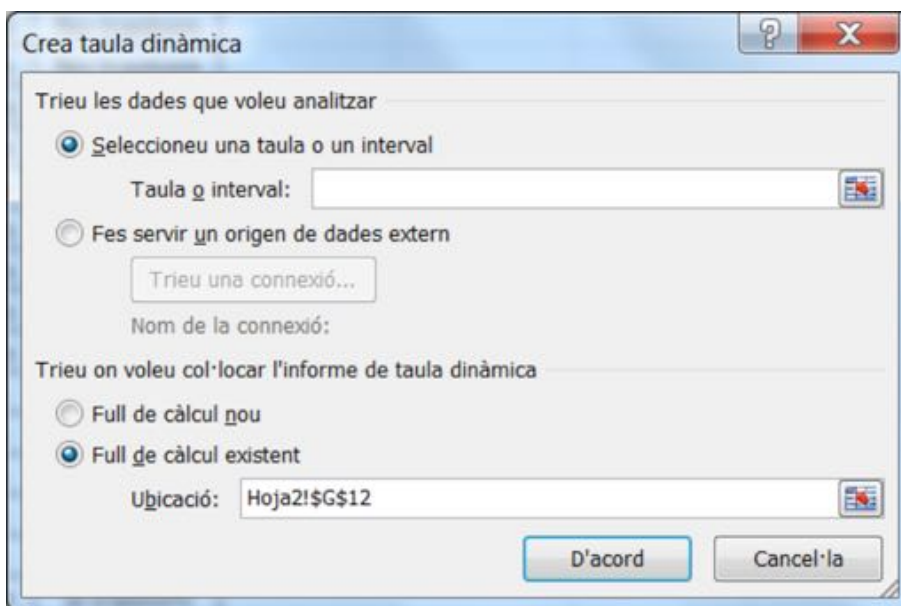
5) La variable us quedarà ordenada de menor a major i ara només caldrà que substituïu els valors igual o inferiors a tres per la categoria «No trastorn» i els valors superiors a tres per «Sí trastorn».

6) Caldrà que repetiu els mateixos passos per a la variable BDI, però ara caldrà que substituïu els valors de la nova variable «BDI_R» de manera que aquells que siguin iguals o inferiors a 15 els correspongui la categoria «No trastorn» i els valors superiors a 15, «Sí trastorn».

7) A partir d'aquestes dades, podreu construir una taula de contingència, mitjançant l'opció de **Taula dinàmica**. Per obtenir-la seguiu la seqüència: **Inserció, Taula dinàmica, Taula dinàmica**.



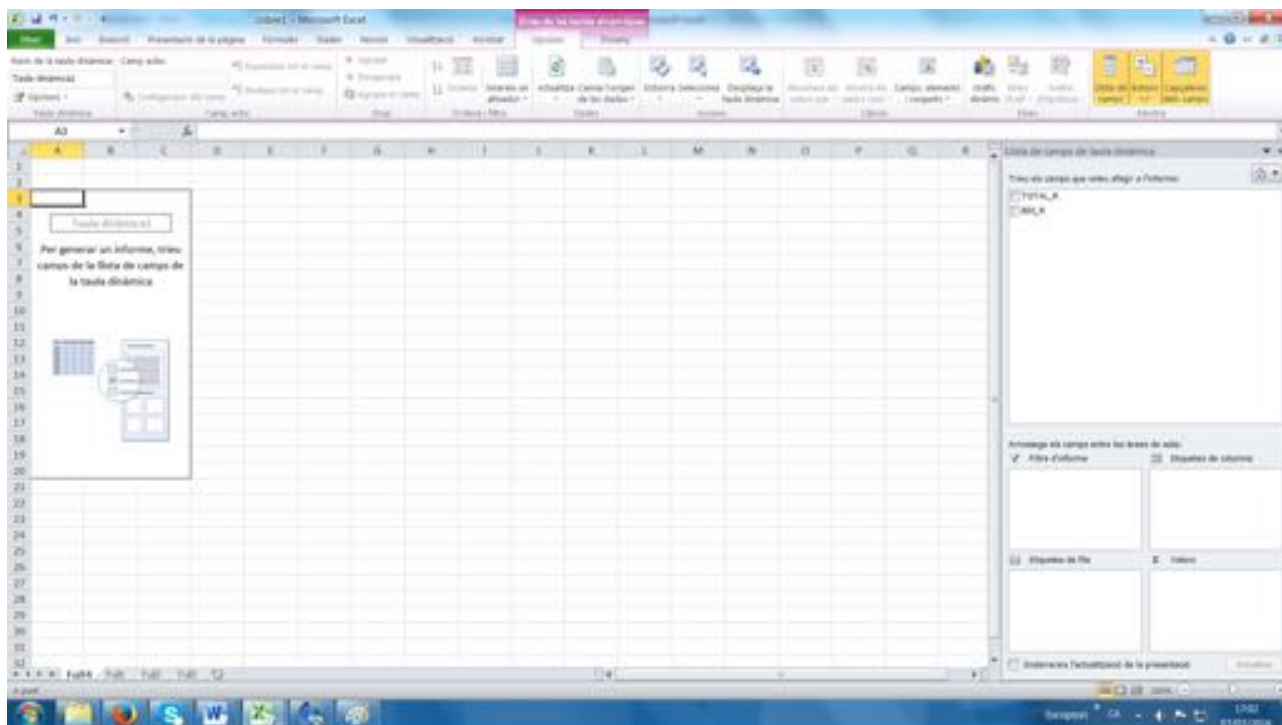
Us apareixerà el quadre de diàleg següent:



8) En el camp **Seleccioneu una taula o un interval**, seleccioneu les caselles on tingueu les dades de les variables TOTAL_R i BDI_R.

9) Marqueu l'opció **Full de càlcul nou**, així la tindreu en un nou full de l'Excel.

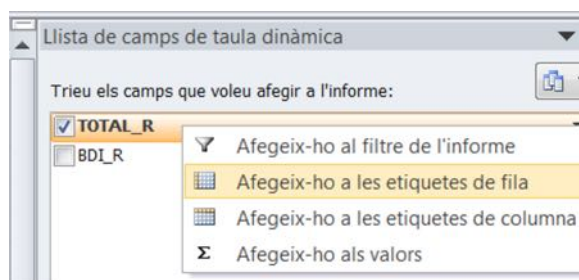
10) Premeu **D'acord** i s'obrirà en un nou full d'Excel la pantalla següent:



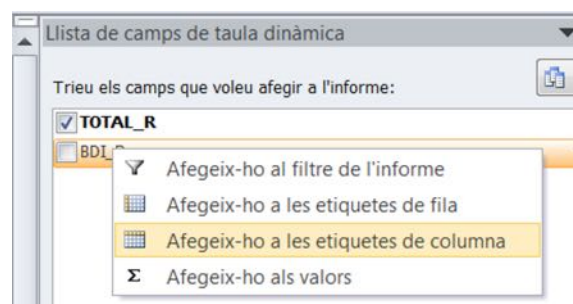
11) A la part esquerra teniu la taula per construir.

12) A la part dreta teniu els camps per omplir:

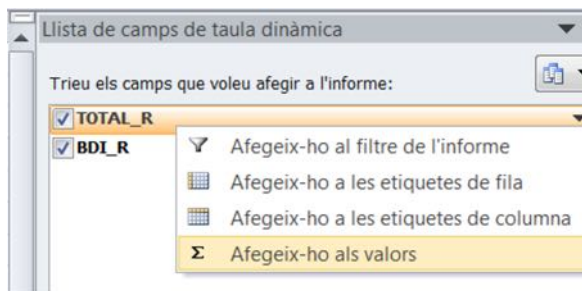
a) Premeu sobre la variable TOTAL_R i clicant amb el botó dret del ratolí seleccioneu l'opció **Afegeix-ho a les etiquetes de fila**.



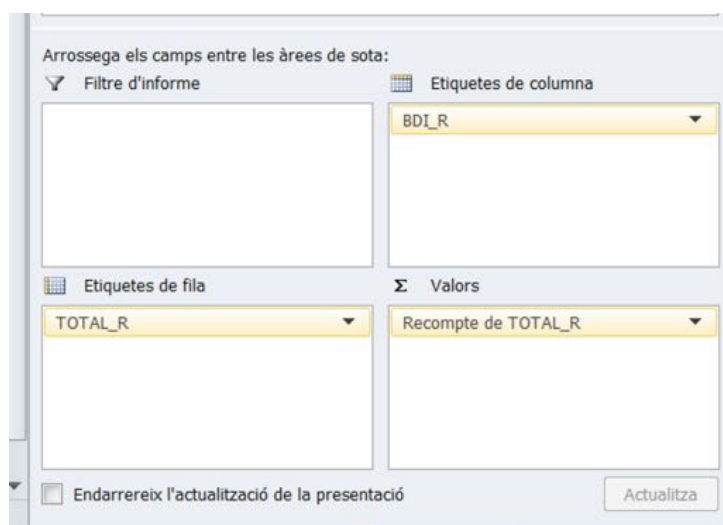
b) Premeu sobre la variable BDI_R i amb el botó dret del ratolí seleccioneu l'opció **Afegeix-ho a les etiquetes de columna**.



c) Premeu sobre qualsevol de les dues variables i amb el botó dret del ratolí seleccioneu **Σ Afegeix-ho als valors**.



La part inferior de la pantalla us ha de quedar de la manera següent:



I a la part dreta de la pantalla us ha d'aparèixer una taula de contingència com la següent:

2					
3	Recompte de TOTAL_R	Etiquetes de columna			
4	Etiquetes de fila	No trastorn	Si trastorn	Total general	
5	No trastorn	14	1	15	
6	Si trastorn	4	6	10	
7	Total general	18	7	25	
8					

A partir dels valors de la taula de contingència podeu calcular fent servir l'Excel com una calculadora els diferents indicadors que es demanen en aquesta pregunta:

- El percentatge d'acord entre els dos tests.
- El coeficient kappa.
- La sensibilitat del nostre test.
- L'especificitat del nostre test.

El **percentatge d'acord** entre els dos tests es calcularia a partir de l'expressió següent:

$$P_c = \frac{14+6}{25} = 0,80$$

Podem interpretar el percentatge d'acord com un nivell d'acord força elevat (80%).

El **coeficient kappa** ve donat per l'expressió següent:

$$k = \frac{20 - 13,6}{25 - 13,6} = 0,56$$

Podem interpretar aquest valor de coeficient kappa com una relació acceptable entre les dues classificacions.

La **sensibilitat del nostre test** la calcularem a partir de l'expressió següent:

$$\text{Sensibilitat} = \frac{\text{Diagnosticats pel nostre test}}{\text{Total diagnosticats (BDI)}} = \frac{6}{7} = 0,86$$

Que representa un 86% de diagnòstics correctes.

L'**especificitat del nostre test** ve donada per la fórmula següent:

$$\text{Especificitat} = \frac{\text{No diagnosticats pel nostre test}}{\text{Total no diagnosticats (BDI)}} = \frac{14}{18} = 0,78$$

Que s'interpreta com un 76% d'encert en la detecció del no trastorn.

3.2. Exercici 2

Per tal de validar l'estructura interna del nostre test, hem fet una anàlisi de components principals (ACP) que ens ha proporcionat una matriu de saturacions factorials, en què s'han extret els dos components amb valors propis superiors a 1.

Component		
Ítem	1	2
Ítem_1	,587	,023
Ítem_2	,267	,793
Ítem_3	,085	,703
Ítem_4	,696	-,303
Ítem_5	,705	-,360

Component		
Ítem	1	2
Ítem_6	,821	,219

A partir d'aquestes dades calculeu les comunalitats dels ítems per a cada component i la conjunta per als dos components, el valor propi de cada component, i la variància explicada per a cada component i la conjunta dels dos components. Interpreteu els resultats respecte a l'estructura interna del test.

Per calcular la comunalitat haureu d'elevat al quadrat les saturacions factorials de cada ítem en cada component. Aquest pas el podem fer a partir de l'Excel fent-lo servir com una calculadora. Per això, en primer lloc cal que introduïu les dades en un full d'Excel tal com estan indicades en el mòdul d'activitats pràctiques.

	A	B	C
1		Components	
2		1	2
3	ítem_1	0,587	0,023
4	ítem_2	0,267	0,793
5	ítem_3	0,085	0,703
6	ítem_4	0,696	-0,303
7	ítem_5	0,705	-0,36
8	ítem_6	0,821	0,219

1) Per calcular la comunalitat de l'ítem 1 en el component 1, heu d'activar l'eina de funcions amb el signe d'igualtat (=).

2) Per calcular el quadrat de la saturació factorial, situeu-vos en la casella D3 i multipliqueu la casella B3 per si mateixa.

	A	B	C	D	E	F
1		Components		Comunalitats		
2		1	2	C1	C2	Conjunta
3	ítem_1	0,587	0,023	=B3*B3		
4	ítem_2	0,267	0,793			
5	ítem_3	0,085	0,703			
6	ítem_4	0,696	-0,303			
7	ítem_5	0,705	-0,36			
8	ítem_6	0,821	0,219			

3) Caldrà repetir aquest pas per cada un dels ítems en el component 1. Utilitzeu l'opció d'autoemplenament per emplenar la resta de caselles. Automàticament apareixeran tots els resultats de les comunalitats de la resta d'ítems.

4) Seguiu el mateix procediment per calcular les comunalitats del segon component.

5) Per calcular la comunalitat conjunta per l'ítem 1, activeu l'eina de funcions amb el signe d'igualtat (=) i sumeu les caselles D3 i E3.

6) Caldrà repetir el pas anterior per cada un dels ítems. Recordeu utilitzar l'opció d'autoemplenament per emplenar la resta de caselles. Automàticament apareixeran tots els resultats de les comunalitats de la resta d'ítems.

	A	B	C	D	E	F
1		Components		Comunalitats		
2		1	2	C1	C2	Conjunta
3	ítem_1	0,587	0,023	0,344569	0,000529	=D3+E3
4	ítem_2	0,267	0,793	0,071289	0,628849	
5	ítem_3	0,085	0,703	0,007225	0,494209	
6	ítem_4	0,696	-0,303	0,484416	0,091809	
7	ítem_5	0,705	-0,36	0,497025	0,1296	
8	ítem_6	0,821	0,219	0,674041	0,047961	

7) Per calcular el valor propi del component 1, feu el sumatori de les comunalitats d'aquest component situant-vos en la casella 9D i fent un doble clic sobre la icona Σ Suma automàtica.



8) Repetiu el pas anterior per calcular el valor propi a partir de les comunalitats del component 2 i de la comunalitat conjunta.

	A	B	C	D	E	F
1		Components		Comunalitats		
2		1	2	C1	C2	Conjunta
3	ítem_1	0,587	0,023	0,344569	0,000529	0,345098
4	ítem_2	0,267	0,793	0,071289	0,628849	0,700138
5	ítem_3	0,085	0,703	0,007225	0,494209	0,501434
6	ítem_4	0,696	-0,303	0,484416	0,091809	0,576225
7	ítem_5	0,705	-0,36	0,497025	0,1296	0,626625
8	ítem_6	0,821	0,219	0,674041	0,047961	0,722002
9	Valor propi			2,078565	1,392957	3,471522

9) Per calcular la variància explicada del primer component caldrà dividir el seu valor propi entre 6 (nombre d'ítems) i multiplicar per 100. Aquest càlcul el podeu fer a partir d'activar les funcions amb el signe = situant-vos a la casella D10, i escriviu l'expressió següent: =(D9/6)*100.

10) Caldrà repetir el mateix procediment per calcular la variància explicada del segon component i del test.

11) La taula us hauria de quedar de la manera següent:

	A	B	C	D	E	F
1		Components		Comunalitats		
2		1	2	C1	C2	Conjunta
3	Ítem_1	0,587	0,023	0,344569	0,000529	0,345098
4	Ítem_2	0,267	0,793	0,071289	0,628849	0,700138
5	Ítem_3	0,085	0,703	0,007225	0,494209	0,501434
6	Ítem_4	0,696	-0,303	0,484416	0,091809	0,576225
7	Ítem_5	0,705	-0,36	0,497025	0,1296	0,626625
8	Ítem_6	0,821	0,219	0,674041	0,047961	0,722002
9	Valor propi			2,078565	1,392957	3,471522
10	Variància explicada			34,64275	23,21595	57,8587

A partir d'aquests resultats podem considerar que el test presenta una estructura interna bidimensional (mesuraria dues dimensions), atès que els diferents ítems presenten saturacions factorials elevades (superiors a 0,30) amb algun dels dos components. Així, els ítems 1, 4, 5 i 6 estarien relacionats amb el component 1, mentre que els ítems 2 i 3 estarien relacionats amb el component 2. El primer component explica el 34,64% de la variabilitat total dels sis ítems, mentre que el segon explica un percentatge del 23,21%. Els dos components extrets conjuntament expliquen un 57,86% de la variabilitat total.

4. Solucions dels exercicis de transformació i interpretació de puntuacions

4.1. Exercici 1

Transformeu les puntuacions directes del test de depressió aplicat en el cas 2 a la mostra de vint-i-cinc subjectes, en les següents puntuacions:

- Percentils.
- Puntuacions estandarditzades.
- Puntuacions T.
- Enneatipus.
- Decatipus.

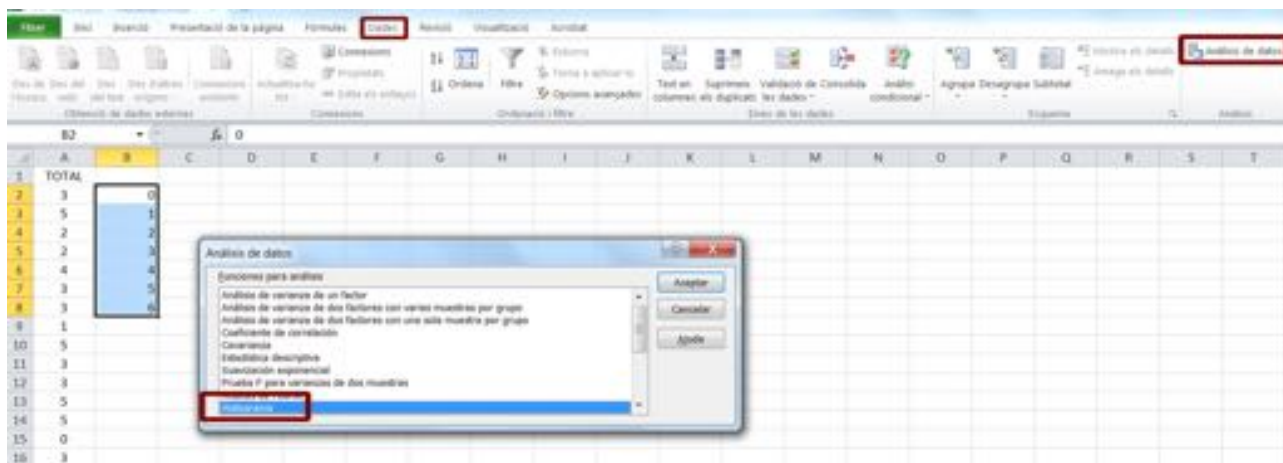
Per fer qualsevol transformació de les puntuacions directes (X) hem de construir la taula de freqüències amb les freqüències absolutes (f_i) per a cada una de les puntuacions en el test. Aquestes freqüències absolutes seran el recompte del nombre de subjectes que han obtingut una puntuació determinada. Podeu obtenir la taula de freqüències amb l'Excel amb el programa preconfigurat d'**Histograma**.

Per fer-ho, heu de copiar les dades de la puntuació total del cas 2 en un full d'Excel, i heu de seguir els passos següents:

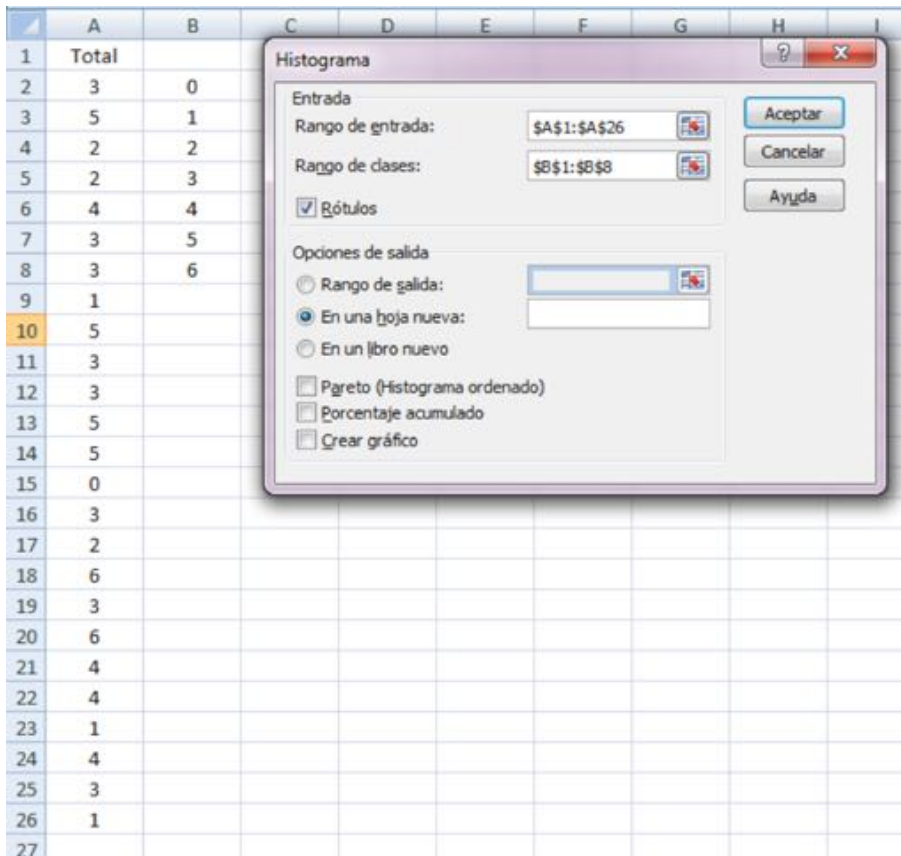
1) Introduïu en unes caselles buides d'una mateixa columna del full els valors de les diferents puntuacions obtingudes en el test. En el nostre cas, aquests valors són 0, 1, 2, 3, 4, 5 i 6. Heu de tenir en compte en introduir aquests valors que heu de deixar una casella buida abans de la primera puntuació. Així, aquestes puntuacions poden ocupar, per exemple, les caselles de B2 a B8.

	A	B
1	Total	
2	3	0
3	5	1
4	2	2
5	2	3
6	4	4
7	3	5
8	3	6
9	1	
10	5	
11	3	
12	3	
13	5	
14	5	
15	0	
16	3	
17	2	
18	6	
19	3	
20	6	
21	4	
22	4	
23	1	
24	4	
25	3	
26	1	
27		

2) Activeu el complement **Herramientas para análisis**, seguint la seqüència:
Dades, Análisis de datos, Histograma.



3) Premeu **Aceptar** i s'obre el quadre de diàleg següent:



4) En el **Rango de entrada** introduïu el rang de les caselles on tingueu les dades de les puntuacions en el test, incloent-hi la primera filera amb l'etiqueta. En el nostre cas, caselles d'A1 a A26.

5) En el **Rango de clases** introduïu el rang de les caselles on tingueu les dades de les puntuacions en el test, incloent-hi la primera casella que heu deixat buida. En el nostre cas, caselles de B1 a B8.

6) Activeu l'opció **Rótulos**.

7) En **Opciones de salida**, mantingueu l'opció per defecte **En una hoja nueva**.

8) Premeu **Aceptar**.

En un nou full us apareixerà la taula de freqüències següent:

	A	B
1		<i>Frecuencia</i>
2	0	1
3	1	3
4	2	3
5	3	8
6	4	4
7	5	4
8	6	2
9	y mayor...	0

Podeu donar format a la taula amb les accions següents:

- Eliminar la filera 9 «y mayor...».
- Posar vores a les caselles.
- Centrar les caselles.
- Etiquetar les puntuacions de la primera columna amb una X.
- Canviar l'etiqueta de la columna B. Posar en cursiva «*fi*» (freqüència absoluta).

	A	B
1	X	<i>fi</i>
2	0	1
3	1	3
4	2	3
5	3	8
6	4	4
7	5	4
8	6	2

Percentils

Per obtenir els percentils necessitem calcular les freqüències acumulades (f_a):

- 1) En la primera columna buida de la dreta, introduïu l'etiqueta de la variable « f_a » en la primera filera.
- 2) En la segona filera d'aquesta mateixa columna, introduïu el valor de la f_i de la primera puntuació. En el nostre cas un 1.
- 3) En la tercera filera, introduïu la fórmula següent: =B3+C2, i premeu **Intro**. Us apareixerà el valor 4 (freqüència acumulada per la segona puntuació).
- 4) Utilitzeu l'opció d'autoemplenament per emplenar la resta de caselles utilitzant la mateixa fórmula. Us apareixeran totes les freqüències acumulades per la resta de puntuacions.

	A	B	C	D
1	X	<i>fi</i>	<i>fa</i>	
2	0	1	1	
3	1	3	4	
4	2	3	7	
5	3	8	15	
6	4	4	19	
7	5	4	23	
8	6	2	25	

A partir d'aquestes dades, per obtenir els percentils (P_c) apliquem la fórmula:

$$P_c = \frac{f_a + 0,5f_i}{N} \times 100$$

on:

- f_a és la freqüència acumulada prèvia a la puntuació directa de la qual es vol calcular el percentil,
- f_i la freqüència absoluta de la puntuació directa, i
- N el nombre de persones que constitueixen la mostra.

Per obtenir els percentils amb l'Excel, seguiu els passos següents:

1) En la primera columna buida de la dreta, introduïu l'etiqueta de la variable P_c en la primera filera.

2) En la segona filera d'aquesta mateixa columna, introduïu la fórmula següent $=((0,5*B2)/25)*100$ i premeu **Intro**. Us apareixerà el valor 2 (percentil per la primera puntuació).

3) En la tercera filera, introduïu la fórmula següent: $=((C2+(0,5*B3))/25)*100$ i premeu **Intro**. Us apareixerà el valor 10 (percentil per la segona puntuació).

4) Utilitzeu l'opció d'autoemplenament per emplenar la resta de caselles utilitzant la mateixa fórmula. Us apareixeran tots els percentils per la resta de puntuacions.

	A	B	C	D	E	F
1	X	f_i	f_a	P_c		
2	0	1	1	2		
3	1	3	4	10		
4	2	3	7	22		
5	3	8	15	44		
6	4	4	19	68		
7	5	4	23	84		
8	6	2	25	96		

Puntuacions estandarditzades

La fórmula de les puntuacions estandarditzades és:

$$Z_x = \frac{X - \bar{X}}{S_x}$$

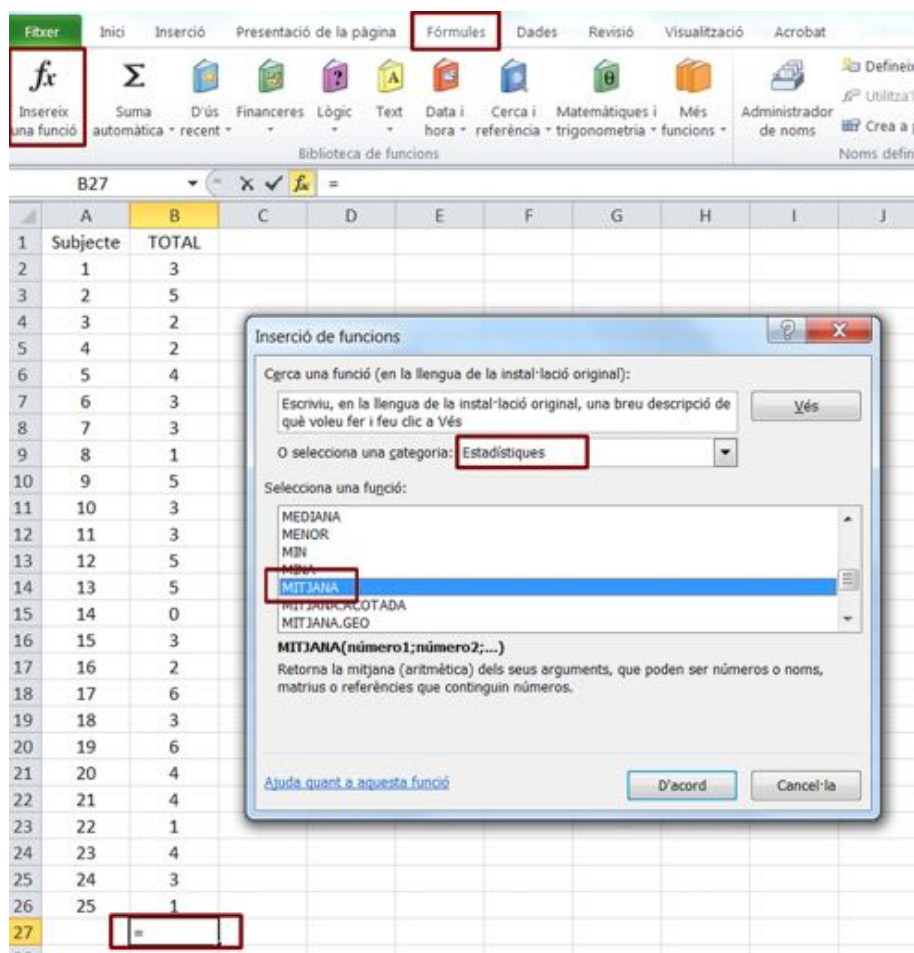
on:

- X és la puntuació directa,
- $\bar{X}$ la mitjana de la mostra, i
- S_x la desviació típica.

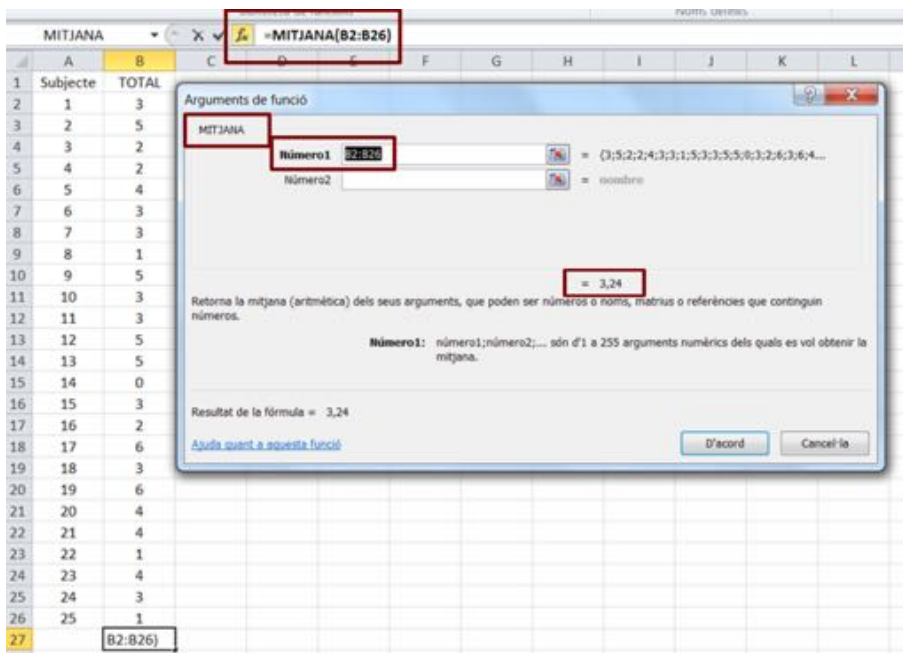
Per tant, hem d'obtenir la mitjana i la desviació típica de la distribució de dades. Podeu fer-ho a partir de les dades sense agrupar, o a partir de les dades agrupades.

Dades sense agrupar:

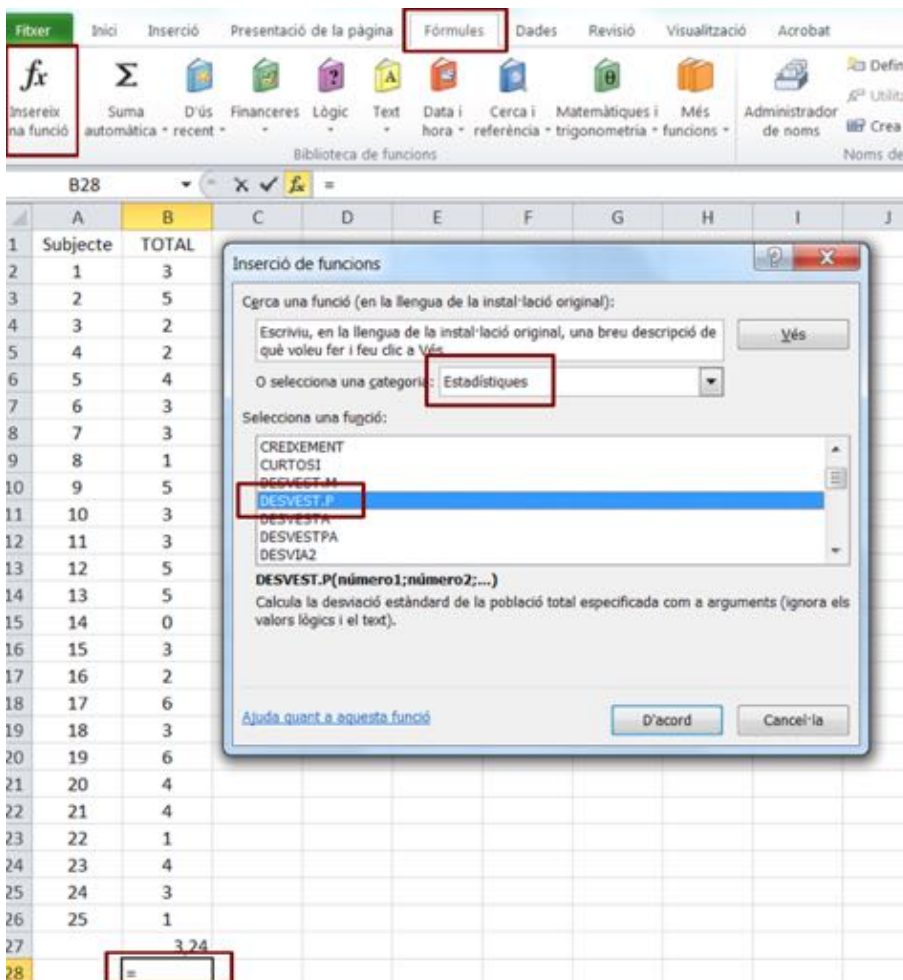
Us situeu a la primera filera lliure de la columna de les puntuacions totals, i seguiu la seqüència: **Fórmules**, **Insereix una funció**, i dins la categoria **Estadístiques**, activeu **Mitjana** i cliqueu **D'acord**.

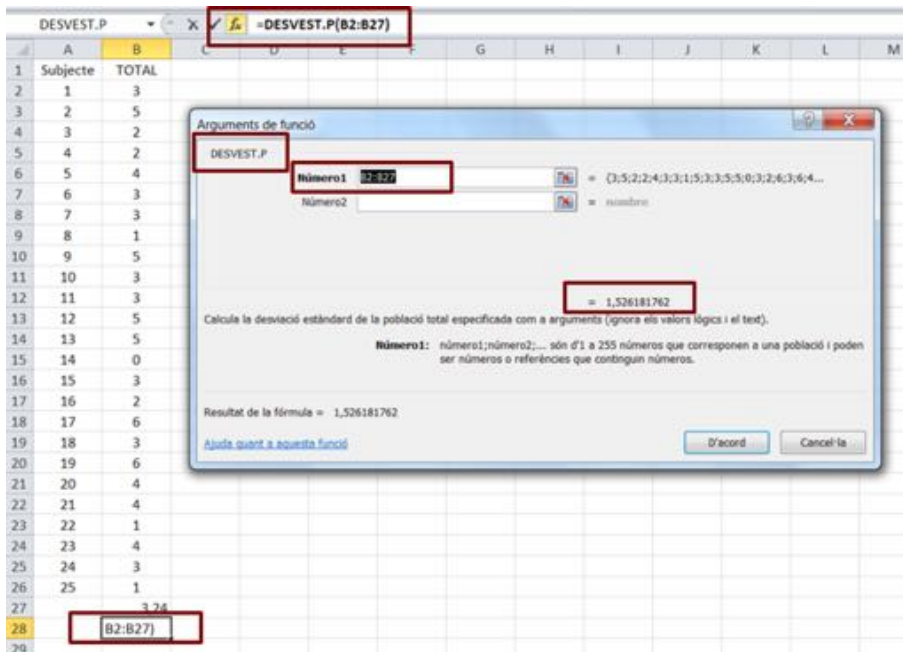


En el quadre de diàleg que se us obre, introduïu a **Número1** el rang de les caselles on teniu les dades de les puntuacions totals dels vint-i-cinc subjectes. En el nostre cas, caselles B2 a B26. Cliqueu **D'acord** i us apareixerà a la casella B27 el valor de la mitjana de les puntuacions totals (3,24).



Per obtenir la desviació típica, seguïu el mateix procediment però amb la funció DESVEST.P.





El valor de la desviació típica arrodonit a dos decimals és d'1,56.

Si les dades estan agrupades, per obtenir la mitjana i la desviació típica amb l'Excel, heu de seguir els passos següents:

1) **Mitjana.** A partir de la taula de freqüències absolutes, heu d'obtenir el producte de la puntuació directa (X) per la seva freqüència (f_i) seguint els passos següents:

a) En la primera columna buida de la dreta, introduïu l'etiqueta de la variable « Xf_i » en la primera filera.

b) En la segona filera d'aquesta mateixa columna, introduïu la fórmula $=A2*B2$ i premeu **Intro**. Us apareixerà el valor 0 (producte de X per f_i de la primera puntuació).

c) Utilitzeu l'opció d'autoemplenament per emplenar la resta de caselles utilitzant la mateixa fórmula. Us apareixeran tots els productes X per f_i per la resta de puntuacions.

	A	B	C	D
1	X	f_i	Xf_i	
2	0	1	0	
3	1	3	3	
4	2	3	6	
5	3	8	24	
6	4	4	16	
7	5	4	20	
8	6	2	12	

e) Us col·loqueu en la primera casella buida d'aquesta tercera columna i introduïu la fórmula $\text{=SUMA}(C2:C8)/25$ i premeu **Intro**. Us apareixerà en aquesta casella el valor 3,24, que és la mitjana de les puntuacions dels vint-i-cinc subjectes.

	A	B	C	D	E
1	X	f_i	Xf_i		
2	0	1	0		
3	1	3	3		
4	2	3	6		
5	3	8	24		
6	4	4	16		
7	5	4	20		
8	6	2	12		
9			3,24		

2) **Desviació típica.** A partir de la taula anterior, heu d'obtenir el producte de la freqüència absoluta f_i pel quadrat de la diferència entre cada puntuació directa X i la mitjana de les puntuacions (3,24).

a) En la primera columna buida de la dreta, introduïu l'etiqueta de la variable « $f_i (X - \text{mitjana})^2$ » en la primera filera.

b) En la segona filera d'aquesta mateixa columna, introduïu la fórmula $\text{=B2}*(A2-3,24)*(A2-3,24)$ i premeu **Intro**. Us apareixerà el valor 10,4976 (producte de f_i pel quadrat de la diferència entre la puntuació directa de 0 i la mitjana de 3,24).

c) Utilitzeu l'opció d'autoemplenament per emplenar la resta de caselles utilitzant la mateixa fórmula. Us apareixeran els productes de f_i pel quadrat de la diferència entre cada puntuació directa i la mitjana per la resta de puntuacions.

	A	B	C	D	E
1	X	f_i	Xf_i	$f_i (X - \text{mitjana})^2$	
2	0	1	0	10,4976	
3	1	3	3	15,0528	
4	2	3	6	4,6128	
5	3	8	24	0,4608	
6	4	4	16	2,3104	
7	5	4	20	12,3904	
8	6	2	12	15,2352	

d) Us col·loqueu en la primera casella buida d'aquesta columna D i introduïu la fórmula $\text{=SUMA}(D2:D8)/25$ i premeu **Intro**. Us apareixerà en aquesta casella el valor 2,4224, que és la variància de les puntuacions dels vint-i-cinc subjectes.

e) Us col·loqueu en la propera casella buida d'aquesta columna D i activeu la funció **ARREL** de la categoria de funcions **Matemàtiques i trigonometria**, i a **Número** introduïu la casella de la variància D9. Premeu **Intro** i us apareix el valor de la desviació típica, 1,56 arrodonit a dos decimals.

The screenshot shows the Excel interface with the **Fórmules** (Formulas) tab selected. The **Insertar una funció** (Insert Function) button is highlighted. The formula bar shows **=ARREL(D9)**. The spreadsheet data is as follows:

	A	B	C	D
1	X	f_i	Xf_i	$f_i (X - \text{mitjana})^2$
2	0	1	0	10,4976
3	1	3	3	15,0528
4	2	3	6	4,6128
5	3	8	24	0,4608
6	4	4	16	2,3104
7	5	4	20	12,3904
8	6	2	12	15,2352
9				2,4224
10				=ARREL(D9)

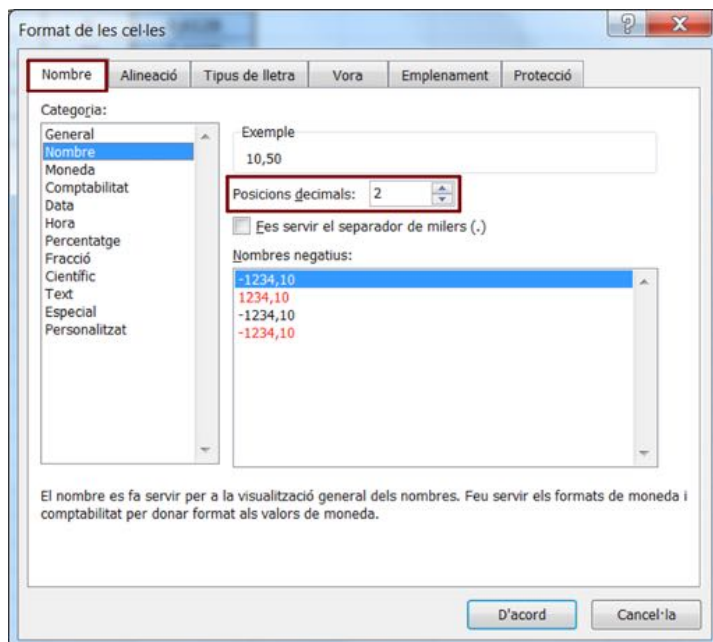
The **Arguments de funció** (Function Arguments) dialog box is open, showing the **ARREL** function. The **Número** (Number) argument is set to **D9**, and the result is displayed as **1,556406117**.

Per arrodonir els valors d'una o diverses caselles a un determinat nombre de decimals, heu d'activar la o les caselles corresponents, i prémer el botó dret del ratolí. En el menú desplegable que s'obre heu d'activar l'opció **Format de cel·les**.

The screenshot shows a right-click context menu over a range of cells in column D. The menu options are:

- Retalla
- Copia
- Opcions d'enganxament:
- Enganxa-ho amb opcions...
- Insereix...
- Suprimeix...
- Suprimeix el contingut
- Filtre
- Ordena
- Insereix un comentari
- Format de les cel·les...**
- Tria a la llista desplegable...
- Defineix el nom...
- Enllaça...

En el quadre de diàleg d'aquesta opció, heu de situar-vos en la finestra de **Número**, i indicar el nombre de decimals que voleu arrodonir a **Posicions decimals**.



Si premeu **D'acord**, els valors de les caselles marcades s'arrodoneixen al nombre de decimals indicat (en el nostre cas, dos):

D
$fi (X - \text{mitjana})^2$
10,50
15,05
4,61
0,46
2,31
12,39
15,24
2,42
1,56

Amb les dades de la mitjana (3,24) i la desviació típica (1,56), ja podeu aplicar la fórmula de la puntuació estandarditzada per a cada puntuació directa.

Amb l'Excel, a partir de la taula de freqüències, seguiu els passos següents:

a) En la primera columna buida de la dreta, introduïu l'etiqueta de la variable « Z_x » en la primera filera.

b) En la segona filera d'aquesta mateixa columna, introduïu la fórmula $=(A2-3,24)/1,56$ i premeu **Intro**. Us apareixerà el valor $-2,08$ (puntuació estandaritzada arrodonida a dos decimals per la puntuació de 0).

c) Utilitzeu l'opció d'autoemplenament per emplenar la resta de caselles utilitzant la mateixa fórmula. Us apareixeran totes les puntuacions estandaritzades per la resta de puntuacions (arrodonides a dos decimals).

	A	B	C	D	E
1	X	<i>f_i</i>	<i>Z_x</i>		
2	0	1	-2,08		
3	1	3	-1,44		
4	2	3	-0,79		
5	3	8	-0,15		
6	4	4	0,49		
7	5	4	1,13		
8	6	2	1,77		

Puntuacions T

A partir de les puntuacions estandaritzades podeu obtenir fàcilment les puntuacions T de McCall a partir de la següent transformació:

$$T = 50 + 10 Z_x$$

Amb l'Excel us situeu en una nova columna (D), introduïu l'etiqueta de la variable en la primera filera (T), i en la segona filera la fórmula $=50+(10*C2)$. Premeu **Intro** i us apareix el valor de 29, que és la puntuació T arrodonida a l'enter més proper (posicions decimals = 0) per la puntuació directa de 0. Per obtenir la resta de puntuacions T, utilitzeu el procediment d'autoemplenament.

	A	B	C	D	E
1	X	<i>f_i</i>	<i>Z_x</i>	T	
2	0	1	-2,08	29,23	
3	1	3	-1,44	35,64	
4	2	3	-0,79	42,05	
5	3	8	-0,15	48,46	
6	4	4	0,49	54,87	
7	5	4	1,13	61,28	
8	6	2	1,77	67,69	
9					

Enneatipus (E)

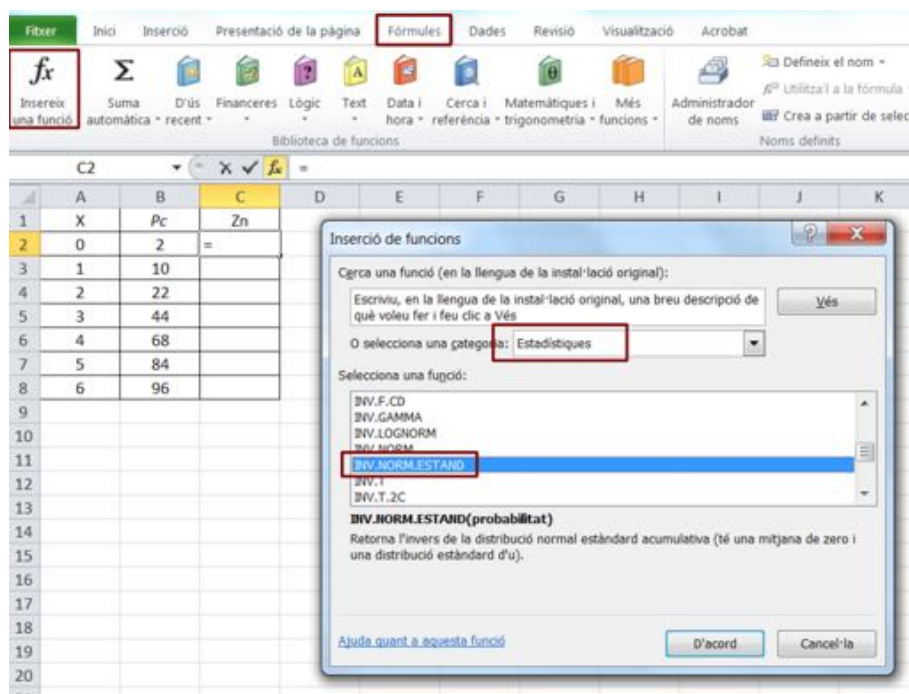
Per obtenir els enneatipus, primer hem de calcular les puntuacions estandaritzades normalitzades (Z_n), transformant els percentils en aquelles puntuacions estandaritzades que sota la taula de la distribució normal tenen associada

una proporció igual al percentil dividit per cent. Per exemple, el percentil 2 que correspon en les nostres dades a la puntuació directa de 0, li correspon una puntuació estandarditzada normalitzada de $-2,05$, que és la puntuació z de la distribució normal associada a una proporció de 0,02 (percentil dividit per cent).

Un cop tenim la Z_n , simplement per obtenir l'enneatipus apliquem la transformació $E = 5 + 2Z_n$, que per la puntuació de 0 ens donarà: $E = 5 + (2 \cdot -2,05) = 0,9$, que arrodonim a l'enter més proper (1), ja que els enneatipus tampoc tenen decimals.

Per obtenir aquests enneatipus (E) amb l'Excel, seguiu els passos següents:

- 1) Recupereu la taula de freqüències, només amb les puntuacions directes (X) i els percentils (P_c).
- 2) Us situeu en la primera filera de la primera columna buida (C1), i introduïu l'etiqueta de la variable «Zn» (puntuacions estandarditzades normalitzades).
- 3) Us col·loqueu en la propera casella buida d'aquesta columna C i activeu la funció **INV.NORM.ESTAND.** de la categoria de funcions **Estadístiques**. Premeu **D'acord**.



- 4) En el quadre de diàleg que us apareix, introduïu a **Probabilitat** la casella de la puntuació típica dividida per 100. Premeu **Intro** i us apareix el valor de $-2,05$, que és la puntuació estandarditzada normalitzada per la puntuació directa de 0, arrodonida a dos decimals.

5) Utilitzeu l'opció d'autoemplenament per emplenar la resta de caselles utilitzant la mateixa fórmula. Us apareixeran totes les puntuacions estandarditzades normalitzades per la resta de puntuacions (arrodonides a dos decimals).

The screenshot shows the Excel interface with the 'Fórmules' ribbon selected. The formula bar displays $=\text{INV.NORM.ESTAND}(B2/100)$. The 'Argumentos de función' dialog box is open, showing the function 'INV.NORM.ESTAND' with the argument 'Probabilidad' set to 'B2/100'. The calculated result is -2,05. The background spreadsheet shows columns A, B, and C with data for X, Pc, and Zn respectively.

	A	B	C
1	X	Pc	Zn
2	0	2	-2,05
3	1	10	-1,28
4	2	22	-0,77
5	3	44	-0,15
6	4	68	0,47
7	5	84	0,99
8	6	96	1,75

Un cop teniu les puntuacions estandarditzades normalitzades, heu d'aplicar la fórmula dels enneatipus: $E = 5 + 2Z_n$ que també s'han d'arrodonir a l'enter més proper.

	D2		=5+(2*C2)	
	A	B	C	D
1	X	Pc	Zn	E
2	0	2	-2,05	1
3	1	10	-1,28	2
4	2	22	-0,77	3
5	3	44	-0,15	5
6	4	68	0,47	6
7	5	84	0,99	7
8	6	96	1,75	9

Decatipus (D)

Els decatipus també són puntuacions normalitzades derivades que s'obtenen de les puntuacions estandarditzades normalitzades a partir de la següent transformació: $D = 5,5 + 2Z_n$. Per tant, amb l'Excel les obtindrem de manera similar als enneatipus però variant la fórmula del càlcul:

E2		$f_x = 5,5 + (2 * C2)$			
	A	B	C	D	E
1	X	Pc	Zn	E	D
2	0	2	-2,05	1	1
3	1	10	-1,28	2	3
4	2	22	-0,77	3	4
5	3	44	-0,15	5	5
6	4	68	0,47	6	6
7	5	84	0,99	7	7
8	6	96	1,75	9	9

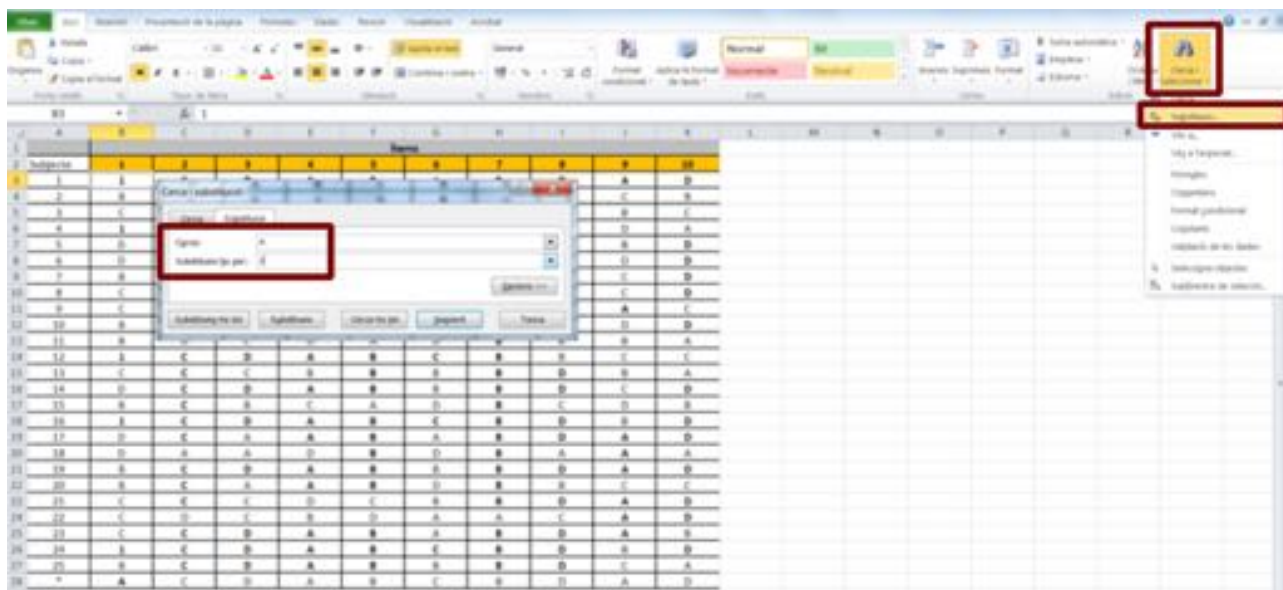
4.2. Exercici 2

En el cas 3 podem considerar que la puntuació total d'un subjecte en el test serà igual al nombre d'ítems que contesta correctament. Per altra banda, també podem considerar que, en ser un test d'intel·ligència, li podem aplicar certes característiques de l'escala Wechler Adult Intelligence Scale (WAIS).

Tenint en compte les consideracions anteriors, transformeu les puntuacions directes dels tres primers subjectes de la matriu de dades del cas 3 en percentils i en coeficient intel·lectual (CI).

Per realitzar el segon exercici, en primer lloc heu de substituir en la matriu de dades original del cas 3 les respostes dels subjectes pels valors 0 o 1 en funció que siguin incorrectes o correctes. Per fer-ho amb l'Excel, seguiu els passos següents:

- 1) Activeu les respostes dels subjectes al primer ítem (caselles de B2 a B26), i seguiu la seqüència següent: **Cerca i selecciona**, **Substitueix**.
- 2) En el quadre de diàleg d'aquesta opció, introduïu una A (resposta correcta de l'ítem 1) a **Cerca**, i un 1 a **Substitueix-ho per**.
- 3) Cliqueu **Substitueix-ho tot** i **D'acord**.



4) Repetiu el mateix procés per l'ítem 2, però posant C (resposta correcta del segon ítem) a **Cerca**, i 1 a **Substitueix-ho per**.

5) Repetiu de nou el mateix procés per als altres ítems, posant sempre a **Cerca** la lletra de la resposta correcta de cada un d'ells, i a **Substitueix-ho per** un 1.

Subjecte	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	A	1	1	1	1
2	B	1	A	1	1	1	1	1	C	B
3	C	1	C	B	A	B	1	A	B	C
4	1	1	1	1	1	1	1	1	D	A
5	D	A	A	B	1	D	1	B	B	1
6	D	B	A	C	C	D	A	A	D	1
7	B	1	C	1	1	A	1	1	C	1
8	C	1	A	1	1	B	1	A	C	1
9	C	1	1	1	1	1	1	1	1	C
10	B	1	A	C	D	A	C	C	D	1
11	B	D	C	D	A	D	1	B	B	A
12	1	1	1	1	1	1	1	B	C	C
13	C	1	C	B	1	B	1	1	B	A
14	D	1	1	1	1	B	1	1	C	1
15	B	1	B	C	A	D	1	C	D	B
16	1	1	1	1	1	1	1	1	B	1
17	D	1	A	1	1	A	1	1	1	1
18	D	A	A	D	1	D	1	A	1	A
19	B	1	1	1	1	B	1	1	1	1
20	B	1	A	1	1	D	1	B	C	C
21	C	1	C	D	C	B	1	1	1	1
22	C	D	C	B	D	A	A	C	1	1
23	C	1	1	1	1	A	1	1	1	B
24	1	1	1	1	1	1	1	1	B	1
25	B	1	1	1	1	B	1	1	C	A

La matriu resultant ja té els valors d'1 corresponents a les respostes correctes de tots els deu ítems. Ara només cal que activeu totes les caselles de les respostes (de B2 a K26), i repetiu el procés anterior posant respectivament A, B, C i D a Cerca i sempre un 0 a **Substitueix-ho per**.

	A	B	C	D	E	F	G	H	I	J	K
1	Subjecte	1	2	3	4	5	6	7	8	9	10
2	1	1	1	1	1	1	0	1	1	1	1
3	2	0	1	0	1	1	1	1	1	0	0
4	3	0	1	0	0	0	0	1	0	0	0
5	4	1	1	1	1	1	1	1	1	0	0
6	5	0	0	0	0	1	0	1	0	0	1
7	6	0	0	0	0	0	0	0	0	0	1
8	7	0	1	0	1	1	0	1	1	0	1
9	8	0	1	0	1	1	0	1	0	0	1
10	9	0	1	1	1	1	1	1	1	1	0
11	10	0	1	0	0	0	0	0	0	0	1
12	11	0	0	0	0	0	0	1	0	0	0
13	12	1	1	1	1	1	1	1	0	0	0
14	13	0	1	0	0	1	0	1	1	0	0
15	14	0	1	1	1	1	0	1	1	0	1
16	15	0	1	0	0	0	0	1	0	0	0
17	16	1	1	1	1	1	1	1	1	0	1
18	17	0	1	0	1	1	0	1	1	1	1
19	18	0	0	0	0	1	0	1	0	1	0
20	19	0	1	1	1	1	0	1	1	1	1
21	20	0	1	0	1	1	0	1	0	0	0
22	21	0	1	0	0	0	0	1	1	1	1
23	22	0	0	0	0	0	0	0	0	1	1
24	23	0	1	1	1	1	0	1	1	1	0
25	24	1	1	1	1	1	1	1	1	0	1
26	25	0	1	1	1	1	0	1	1	0	0

Finalment, cal que calculeu la puntuació total de cada subjecte, que serà el recompte de les seves respostes correctes, o sigui el sumatori dels valors dels deu ítems. Per fer-ho amb l'Excel, seguiu els passos següents:

- 1) En la primera columna buida de la dreta (L), introduïu l'etiqueta de la variable X en la primera filera.
- 2) En la segona filera d'aquesta mateixa columna (L2), introduïu la fórmula =SUMA(B2:K2) i premeu **Intro**. Us apareix el valor de 9, que és la puntuació directa del primer subjecte.
- 3) Utilitzeu l'opció d'autoemplenament per emplenar la resta de caselles utilitzant la mateixa fórmula. Us apareixeran totes puntuacions directes per la resta de subjectes.

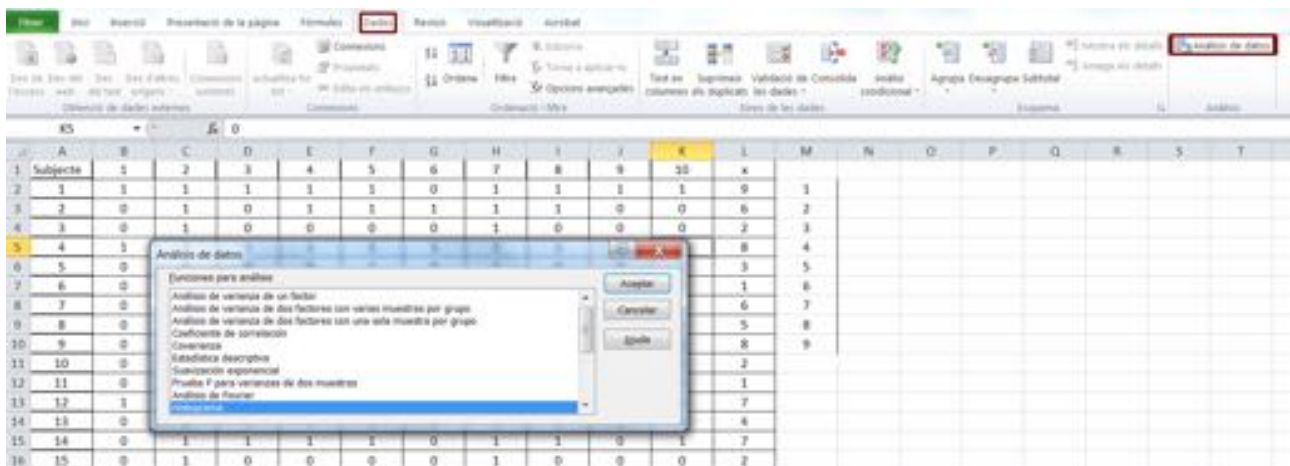
	A	B	C	D	E	F	G	H	I	J	K	L
1	Subjecte	1	2	3	4	5	6	7	8	9	10	x
2	1	1	1	1	1	1	0	1	1	1		=SUMA(B2:K2)
3	2	0	1	0	1	1	1	1	1	0	0	6
4	3	0	1	0	0	0	0	1	0	0	0	2
5	4	1	1	1	1	1	1	1	1	0	0	8
6	5	0	0	0	0	1	0	1	0	0	1	3
7	6	0	0	0	0	0	0	0	0	0	1	1
8	7	0	1	0	1	1	0	1	1	0	1	6
9	8	0	1	0	1	1	0	1	0	0	1	5
10	9	0	1	1	1	1	1	1	1	1	0	8
11	10	0	1	0	0	0	0	0	0	0	1	2
12	11	0	0	0	0	0	0	1	0	0	0	1
13	12	1	1	1	1	1	1	1	0	0	0	7
14	13	0	1	0	0	1	0	1	1	0	0	4
15	14	0	1	1	1	1	0	1	1	0	1	7
16	15	0	1	0	0	0	0	1	0	0	0	2
17	16	1	1	1	1	1	1	1	1	0	1	9
18	17	0	1	0	1	1	0	1	1	1	1	7
19	18	0	0	0	0	1	0	1	0	1	0	3
20	19	0	1	1	1	1	0	1	1	1	1	8
21	20	0	1	0	1	1	0	1	0	0	0	4
22	21	0	1	0	0	0	0	1	1	1	1	5
23	22	0	0	0	0	0	0	0	0	1	1	2
24	23	0	1	1	1	1	0	1	1	1	0	7
25	24	1	1	1	1	1	1	1	1	0	1	9
26	25	0	1	1	1	1	0	1	1	0	0	6
27												

Per calcular els percentils corresponents a les puntuacions directes heu d'obtenir la taula de freqüències. Amb l'Excel ho podeu fer seguint els passos següents:

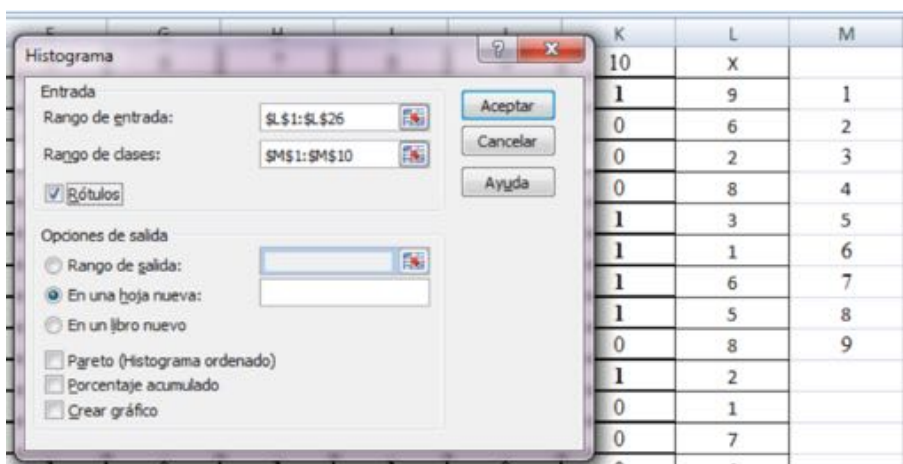
1) Introduïu en unes caselles buides d'una mateixa columna del full els valors de les diferents puntuacions obtingudes en el test. En aquest cas, aquests valors són 1, 2, 3, 4, 5, 6, 7, 8 i 9. Heu de tenir en compte en introduir aquests valors que heu de deixar una casella buida abans de la primera puntuació. Així, aquestes puntuacions poden ocupar, per exemple, les caselles M2 a M10.

L	M
X	
9	1
6	2
2	3
8	4
3	5
1	6
6	7
5	8
8	9
2	
1	

2) Activeu el programa preconfigurat, seguint la seqüència: **Dades, Anàlisis de datos, Histograma**.



3) Premeu **Aceptar** i s'obre el quadre de diàleg següent:



4) En el **Rango de entrada** introduïu el rang de les caselles on tingueu les dades de les puntuacions en el test, incloent-hi la primera filera amb l'etiqueta. En aquest cas, caselles de L1 a L26.

5) En el **Rango de clases** introduïu el rang de les caselles on tingueu les dades de les puntuacions en el test, incloent-hi la primera casella que heu deixat buida. En aquest cas, caselles de M1 a M10.

6) Activeu l'opció **Rótulos**.

7) En **Opciones de salida**, mantingueu l'opció per defecte **En una hoja nueva**.

8) Premeu **Aceptar**.

En un nou full us apareixerà la taula de freqüències següent:

	A	B
1		<i>Frecuencia</i>
2	1	2
3	2	4
4	3	2
5	4	2
6	5	2
7	6	3
8	7	4
9	8	3
10	9	3
11	y mayor...	0

Podeu donar format a la taula amb les accions següents:

- Eliminar la filera 11 «y mayor...».
- Posar vores a les caselles.
- Centrar les caselles.
- Etiquetar les puntuacions de la primera columna amb una X.
- Canviar l'etiqueta de la columna B. Posar en cursiva «*fi*» (freqüència absoluta).

	A	B
1	X	<i>fi</i>
2	1	2
3	2	4
4	3	2
5	4	2
6	5	2
7	6	3
8	7	4
9	8	3
10	9	3

Percentils

Per obtenir els percentils dels tres primers subjectes, necessitem calcular les freqüències acumulades (f_a):

- 1) En la primera columna buida de la dreta, introduïu l'etiqueta de la variable « f_a » en la primera filera.
- 2) En la segona filera d'aquesta mateixa columna, introduïu el valor de la f_i de la primera puntuació. En el nostre cas un 2.
- 3) En la tercera filera, introduïu la fórmula =B3+C2 i premeu **Intro**. Us apareixerà el valor 6 (freqüència acumulada per la segona puntuació).

4) Utilitzeu l'opció d'autoemplenament per emplenar la resta de caselles utilitzant la mateixa fórmula. Us apareixeran totes les freqüències acumulades per la resta de puntuacions.

	A	B	C	D
1	X	f_i	f_a	
2	1	2	2	
3	2	4	6	
4	3	2	8	
5	4	2	10	
6	5	2	12	
7	6	3	15	
8	7	4	19	
9	8	3	22	
10	9	3	25	
11				

A partir d'aquestes dades, per obtenir els percentils (P_c) apliquem la fórmula:

$$P_c = \frac{f_a + 0,5f_i}{N} \times 100$$

on:

- f_a és la freqüència acumulada prèvia a la puntuació directa de la qual es vol calcular el percentil,
- f_i la freqüència absoluta de la puntuació directa, i
- N el nombre de persones que constitueixen la mostra.

Per obtenir els percentils amb l'Excel, seguiu els passos següents:

1) En la primera columna buida de la dreta, introduïu l'etiqueta de la variable « P_c » en la primera filera.

2) En la segona filera d'aquesta mateixa columna, introduïu la fórmula $=((0,5*B2)/25)*100$ i premeu **Intro**. Us apareixerà el valor 4 (percentil per la primera puntuació).

3) En la tercera filera, introduïu la fórmula $=((C2+(0,5*B3))/25)*100$ i premeu **Intro**. Us apareixerà el valor 16 (percentil per la segona puntuació).

4) Utilitzeu l'opció d'autoemplenament per emplenar la resta de caselles utilitzant la mateixa fórmula. Us apareixeran tots els percentils per la resta de puntuacions.

D3					
	A	B	C	D	E
1	X	<i>fi</i>	<i>fa</i>	<i>Pc</i>	
2	1	2	2	4	
3	2	4	6	16	
4	3	2	8	28	
5	4	2	10	36	
6	5	2	12	44	
7	6	3	15	54	
8	7	4	19	68	
9	8	3	22	82	
10	9	3	25	94	
11					

Així, els percentils dels tres primers subjectes són:

- Primer subjecte: X = 9. Percentil = 94
- Segon subjecte: X = 6. Percentil = 54
- Segon subjecte: X = 2. Percentil = 16

Finalment, per transformar les puntuacions directes en *CI*, hem d'obtenir en primer lloc la puntuació estandarditzada *z*, i en segon lloc transformar aquesta puntuació *z* en *CI*.

La fórmula de la puntuació *z* és:

$$Z_x = \frac{X - \bar{X}}{S_x}$$

que requereix el càlcul previ de la mitjana i la desviació típica.

1) Mitjana. A partir de la taula de freqüències absolutes, heu d'obtenir el producte de la puntuació directa (X) per la seva freqüència (*fi*) seguint els passos següents:

- En la primera columna buida de la dreta, introduïu l'etiqueta de la variable «X*fi*» en la primera filera.
- En la segona filera d'aquesta mateixa columna, introduïu la fórmula =A2*B2 i premeu **Intro**. Us apareixerà el valor 2 (producte d'X per *fi* de la primera puntuació).
- Utilitzeu l'opció d'autoemplenament per emplenar la resta de caselles utilitzant la mateixa fórmula. Us apareixeran tots els productes X per *fi* per la resta de puntuacions.

	A	B	C	D
1	X	<i>fi</i>	<i>Xfi</i>	
2	1	2	2	
3	2	4	8	
4	3	2	6	
5	4	2	8	
6	5	2	10	
7	6	3	18	
8	7	4	28	
9	8	3	24	
10	9	3	27	

d) Us col·loqueu en la primera casella buida d'aquesta tercera columna i introduïu la fórmula =SUMA(C2:C10)/25 i premeu **Intro**. Us apareixerà en aquesta casella el valor 5,24, que és la mitjana de les puntuacions dels vint-i-cinc subjectes.

	A	B	C	D	E
1	X	<i>fi</i>	<i>Xfi</i>		
2	1	2	2		
3	2	4	8		
4	3	2	6		
5	4	2	8		
6	5	2	10		
7	6	3	18		
8	7	4	28		
9	8	3	24		
10	9	3	27		
11			5,24		

2) **Desviació típica.** A partir de la taula anterior, heu d'obtenir el producte de la freqüència absoluta *fi* pel quadrat de la diferència entre cada puntuació directa X i la mitjana de les puntuacions (5,24).

a) En la primera columna buida de la dreta, introduïu l'etiqueta de la variable «*fi (X-mitjana)²*» en la primera filera.

b) En la segona filera d'aquesta mateixa columna, introduïu la fórmula =B2*(A2-5,24)*(A2-5,24) i premeu **Intro**. Us apareixerà el valor 35,9552 (producte de *fi* pel quadrat de la diferència entre la puntuació directa d'1 i la mitjana de 5,24).

c) Utilitzeu l'opció d'autoemplenament per emplenar la resta de caselles utilitzant la mateixa fórmula. Us apareixeran els productes de *fi* pel quadrat de la diferència entre cada puntuació directa i la mitjana per la resta de puntuacions.

	D2			
	A	B	C	D
1	X	f_i	Xf_i	$f_i (X-\text{mitjana})^2$
2	1	2	2	35,9552
3	2	4	8	41,9904
4	3	2	6	10,0352
5	4	2	8	3,0752
6	5	2	10	0,1152
7	6	3	18	1,7328
8	7	4	28	12,3904
9	8	3	24	22,8528
10	9	3	27	42,4128

d) Us col·loqueu en la primera casella buida d'aquesta columna D i introduïu la fórmula $\text{=SUMA}(D2:D10)/25$ i premeu **Intro**. Us apareixerà en aquesta casella el valor 6,8224, que és la variància de les puntuacions dels vint-i-cinc subjectes.

e) Us col·loqueu en la propera casella buida d'aquesta columna D i activeu la funció **ARREL** de la categoria de funcions **Matemàtiques i trigonometria**, i a **Número** introduïu la casella de la variància D11. Premeu **Intro** i us apareix el valor de la desviació típica 2,61 arrodonit a dos decimals.

The screenshot shows the Excel interface with the 'Fórmules' ribbon selected. The 'Inserir funció' button is highlighted. The formula bar displays $\text{=ARREL}(D11)$. The 'Arguments de funció' dialog box is open, showing the 'Número' field with the reference D11, which corresponds to the variance value 6,8224. The calculated standard deviation is shown as 2,611972435. The spreadsheet data is as follows:

	A	B	C	D
1	X	f_i	Xf_i	$f_i (X-\text{mitjana})^2$
2	1	2	2	35,9552
3	2	4	8	41,9904
4	3	2	6	10,0352
5	4	2	8	3,0752
6	5	2	10	0,1152
7	6	3	18	1,7328
8	7	4	28	12,3904
9	8	3	24	22,8528
10	9	3	27	42,4128
11				6,8224
12				2,61

Amb les dades de la mitjana (5,24) i la desviació típica (2,61), ja podeu aplicar la fórmula de la puntuació estandarditzada per cada puntuació directa.

Amb l'Excel, a partir de la taula de freqüències, seguïu els passos següents:

1) En la primera columna buida de la dreta, introduïu l'etiqueta de la variable « Z_x » en la primera filera.

2) En la segona filera d'aquesta mateixa columna, introduïu la fórmula $=(A2-3,24)/1,56$ i premeu **Intro**. Us apareixerà el valor $-1,62$ (puntuació estandarditzada arrodonida a dos decimals per la puntuació d'1).

3) Utilitzeu l'opció d'autoemplenament per emplenar la resta de caselles utilitzant la mateixa fórmula. Us apareixeran totes les puntuacions estandarditzades per la resta de puntuacions (arrodonides a dos decimals).

	E2			f_x	$=(A2-5,24)/2,61$
	A	B	C	D	E
1	X	f_i	Xf_i	$f_i (X-\text{mitjana})^2$	Zx
2	1	2	2	35,9552	-1,62
3	2	4	8	41,9904	-1,24
4	3	2	6	10,0352	-0,86
5	4	2	8	3,0752	-0,48
6	5	2	10	0,1152	-0,09
7	6	3	18	1,7328	0,29
8	7	4	28	12,3904	0,67
9	8	3	24	22,8528	1,06
10	9	3	27	42,4128	1,44

Així, les puntuacions estandarditzades dels tres primers subjectes són les següents:

$$X_1 = 9 \quad X_2 = 6 \quad X_3 = 2$$

Amb aquestes puntuacions ja podem obtenir els valors dels coeficients d'intel·ligència (CI). La transformació en CI segueix la següent expressió:

$$CI = 100 + (z \cdot 15)$$

Per fer-ho amb l'Excel, seguiu els passos següents:

1) En la primera columna buida de la dreta, introduïu l'etiqueta de la variable «CI» en la primera filera.

2) En la segona filera d'aquesta mateixa columna, introduïu la fórmula $=100+(E2*15)$ i premeu **Intro**. Us apareixerà el valor 76 (puntuació de CI per la puntuació d'1 arrodonida a l'enter més proper).

3) Utilitzeu l'opció d'autoemplenament per emplenar la resta de caselles utilitzant la mateixa fórmula. Us apareixeran totes les puntuacions de CI per la resta de puntuacions (arrodonides a l'enter més proper).

F2 $f_x = 100 + (E2 * 15)$						
	A	B	C	D	E	F
1	X	f_i	Xf_i	$f_i (X - \text{mitjana})^2$	Zx	CI
2	1	2	2	35,9552	-1,62	76
3	2	4	8	41,9904	-1,24	81
4	3	2	6	10,0352	-0,86	87
5	4	2	8	3,0752	-0,48	93
6	5	2	10	0,1152	-0,09	99
7	6	3	18	1,7328	0,29	104
8	7	4	28	12,3904	0,67	110
9	8	3	24	22,8528	1,06	116
10	9	3	27	42,4128	1,44	122

Així, podem resumir les transformacions pels tres primers subjectes de la matriu de dades en la següent taula:

Subjecte	Puntuació directa	Percentil	CI
1	9	94	122
2	6	54	104
3	2	16	81

5. Solucions dels exercicis d'anàlisi d'ítems

5.1. Exercici 1

A partir de les dades del cas 3, calculeu i interpreteu els índexs de dificultat i de discriminació dels tres primers ítems del test. Per tal d'obtenir els índexs de discriminació considereu com a grup d'alt rendiment o alt nivell d'intel·ligència els subjectes que encerten més de set ítems, i de baix rendiment o baix nivell d'intel·ligència els que encerten menys de tres ítems. La grandària d'aquests dos grups no arriba al 27% o 25% del total dels vint-i-cinc subjectes, que és el recomanat, però són molt a prop d'aquests percentatges.

En primer lloc caldrà classificar els subjectes en els diferents grups de rendiment (alt, mitjà i baix). Per això haurem de calcular la puntuació total de cada subjecte en el test. Com que els valors dels ítems estan expressats en lletres haurem de transformar cada lletra en un valor numèric: 1 per l'encert i 0 per l'error. Una manera relativament ràpida per fer aquesta transformació de lletres a números és utilitzar filtres i substituir un determinat valor per un altre. A continuació es mostra com realitzar aquesta transformació a partir dels valors del cas 3:

- 1) Per fer la transformació de lletres a números, copieu la matriu en un full nou (més endavant necessitarem també els valors de les lletres).
- 2) Per transformar els encerts del primer ítem (lletra A) en 1, seleccioneu amb el ratolí la primera filera on es troba el número d'ítem i activarem l'opció **Filtre** del bloc de **Dades**.

	A	B	C	D	E	F	G	H	I	J	K
1											
2	Subjecte	1	2	3	4	5	6	7	8	9	10
3	1	A	C	D	A	B	A	B	D	A	D
4	2	B	C	A	A	B	C	B	D	C	B
5	3	C	C	C	B	A	B	B	A	B	C
6	4	A	C	D	A	B	C	B	D	D	A
7	5	D	A	A	B	B	D	B	B	B	D
8	6	D	B	A	C	C	D	A	A	D	D
9	7	B	C	C	A	B	A	B	D	C	D
10	8	C	C	A	A	B	B	B	A	C	D
11	9	C	C	D	A	B	C	B	D	A	C
12	10	B	C	A	C	D	A	C	C	D	D
13	11	B	D	C	D	A	D	B	B	B	A
14	12	A	C	D	A	B	C	B	B	C	C
15	13	C	C	C	B	B	B	B	D	B	A
16	14	D	C	D	A	B	B	B	D	C	D
17	15	B	C	B	C	A	D	B	C	D	B
18	16	A	C	D	A	B	C	B	D	B	D
19	17	D	C	A	A	B	A	B	D	A	D
20	18	D	A	A	D	B	D	B	A	A	A
21	19	B	C	D	A	B	B	B	D	A	D
22	20	B	C	A	A	B	D	B	B	C	C
23	21	C	C	C	D	C	B	B	D	A	D
24	22	C	D	C	B	D	A	A	C	A	D
25	23	C	C	D	A	B	A	B	D	A	B
26	24	A	C	D	A	B	C	B	D	B	D
27	25	B	C	D	A	B	B	B	D	C	A
28	*	A	C	D	A	B	C	B	D	A	D

3) Un cop activat el filtre, caldrà prémer la icona que apareix a la capçalera dels noms dels ítems i apareixeran diferents opcions de filtre.

4) Seleccioneu només la resposta A (l'encert).

	A	B	C	D	E	F	G	H	I	J	K
1											
2	Subjecte	1	2	3	4	5	6	7	8	9	10
3	Orderna de l'A a la Z			D	A	B	A	B	D	A	D
4	Orderna de la Z a l'A			A	A	B	C	B	D	C	B
5	Orderna per color			C	B	A	B	B	A	B	C
6	Suprimeix el filtre de "1"			D	A	B	C	B	D	D	A
7	Filtra per color			A	B	B	D	B	B	B	D
8	Filtres de text			A	C	C	D	A	A	D	D
9	Cerca			C	A	B	A	B	D	C	D
10	(Selecció-ho tot)			A	A	B	B	B	A	C	D
11	☒ A			D	A	B	C	B	D	A	C
12	☐ B			A	C	D	A	C	C	D	D
13	☐ C			C	D	A	D	B	B	B	A
14	☐ D			D	A	B	C	B	B	C	C
15	D'acord			C	B	B	B	B	D	B	A
16	Cancel·la			D	A	B	B	B	D	C	D
17				B	C	A	D	B	C	D	B
18				D	A	B	C	B	D	B	D
19				A	A	B	A	B	D	A	D
20				A	D	B	D	B	A	A	A
21	19	B	C	D	A	B	B	B	D	A	D
22	20	B	C	A	A	B	D	B	B	C	C
23	21	C	C	C	D	C	B	B	D	A	D
24	22	C	D	C	B	D	A	A	C	A	D
25	23	C	C	D	A	B	A	B	D	A	B
26	24	A	C	D	A	B	C	B	D	B	D
27	25	B	C	D	A	B	B	B	D	C	A
28	*	A	C	D	A	B	C	B	D	A	D

5) Un cop tingueu seleccionada la resposta A, només us apareixeran aquestes respostes:

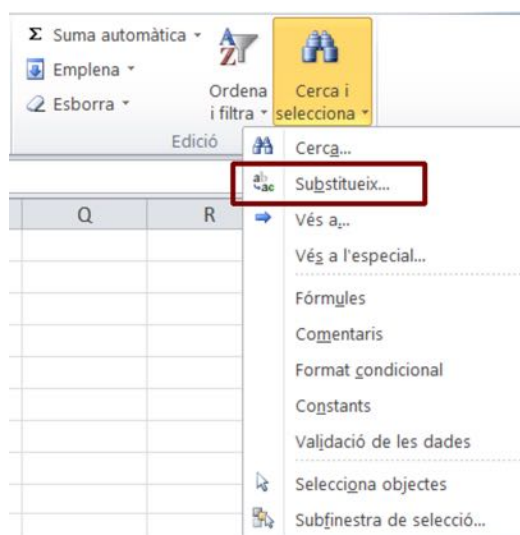
	A	B	C	D	E	F	G	H	I	J	K
1											
2	Subjecte	1	2	3	4	5	6	7	8	9	10
3	1	A	C	D	A	B	A	B	D	A	D
6	4	A	C	D	A	B	C	B	D	D	A
14	12	A	C	D	A	B	C	B	B	C	C
18	16	A	C	D	A	B	C	B	D	B	D
26	24	A	C	D	A	B	C	B	D	B	D
28	*	A	C	D	A	B	C	B	D	A	D

6) Substitueix la primera resposta A de l'ítem 1 per un 1 (per canviar la resta podeu utilitzar l'opció d'autoemplenament).

	A	B	C	D	E	F	G	H	I	J	K
1	Ítems										
2	Subjecte	1	2	3	4	5	6	7	8	9	10
3	1	1	C	D	A	B	A	B	D	A	D
6	4	1	C	D	A	B	C	B	D	D	A
14	12	1	C	D	A	B	C	B	B	C	C
18	16	1	C	D	A	B	C	B	D	B	D
26	24	1	C	D	A	B	C	B	D	B	D
28	*	1	C	D	A	B	C	B	D	A	D

7) Seguiu els mateixos passos per transformar la lletra de l'opció correcta de la resta d'ítems del test en un número.

8) Un cop hagueu substituït tots els encerts per 1 en tots els ítems, premeu el bloc de **Cerca i selecciona** i seleccioneu l'opció **Substitueix**.



9) Marqueu l'opció **Substitució**.

10) Indiqueu en el camp **Cerca** la lletra A.

11) Introduïu el valor 0 en **Substitueix-ho per**.

12) Premeu **Substitueix-ho tot**.

13) Repetiu aquests passos per canviar la resta de lletres (B, C i D).

La base de dades ara tindrà l'aspecte següent:

	A	B	C	D	E	F	G	H	I	J	K
1		Ítems									
2		1	2	3	4	5	6	7	8	9	10
3	1	1	1	1	1	1	0	1	1	1	1
4	2	0	1	0	1	1	1	1	1	0	0
5	3	0	1	0	0	0	0	1	0	0	0
6	4	1	1	1	1	1	1	1	1	0	0
7	5	0	0	0	0	1	0	1	0	0	1
8	6	0	0	0	0	0	0	0	0	0	1
9	7	0	1	0	1	1	0	1	1	0	1
10	8	0	1	0	1	1	0	1	0	0	1
11	9	0	1	1	1	1	1	1	1	1	0
12	10	0	1	0	0	0	0	0	0	0	1
13	11	0	0	0	0	0	0	1	0	0	0
14	12	1	1	1	1	1	1	1	0	0	0
15	13	0	1	0	0	1	0	1	1	0	0
16	14	0	1	1	1	1	0	1	1	0	1
17	15	0	1	0	0	0	0	1	0	0	0
18	16	1	1	1	1	1	1	1	1	0	1
19	17	0	1	0	1	1	0	1	1	1	1
20	18	0	0	0	0	1	0	1	0	1	0
21	19	0	1	1	1	1	0	1	1	1	1
22	20	0	1	0	1	1	0	1	0	0	0
23	21	0	1	0	0	0	0	1	1	1	1
24	22	0	0	0	0	0	0	0	0	1	1
25	23	0	1	1	1	1	0	1	1	1	0
26	24	1	1	1	1	1	1	1	1	0	1
27	25	0	1	1	1	1	0	1	1	0	0

Un cop finalitzada la transformació de lletres a números podreu calcular la puntuació del test per cada subjecte seguint els passos següents:

- 1) Feu el sumatori dels deu ítems per al subjecte 1.
- 2) A partir de l'opció d'autoemplenar, calculeu el sumatori per la resta dels vint-i-cinc subjectes.

	A	B	C	D	E	F	G	H	I	J	K	L
1												
2												
3	1	1	1	1	1	1	0	1	1	1	1	=suma(B3:K3)
4	2	0	1	0	1	1	1	1	1	0	0	
5	3	0	1	0	0	0	0	1	0	0	0	
6	4	1	1	1	1	1	1	1	1	0	0	
7	5	0	0	0	0	1	0	1	0	0	1	
8	6	0	0	0	0	0	0	0	0	0	1	
9	7	0	1	0	1	1	0	1	1	0	1	
10	8	0	1	0	1	1	0	1	0	0	1	
11	9	0	1	1	1	1	1	1	1	1	0	
12	10	0	1	0	0	0	0	0	0	0	1	
13	11	0	0	0	0	0	0	1	0	0	0	
14	12	1	1	1	1	1	1	1	0	0	0	
15	13	0	1	0	0	1	0	1	1	0	0	
16	14	0	1	1	1	1	0	1	1	0	1	
17	15	0	1	0	0	0	0	1	0	0	0	
18	16	1	1	1	1	1	1	1	1	0	1	
19	17	0	1	0	1	1	0	1	1	1	1	
20	18	0	0	0	0	1	0	1	0	1	0	
21	19	0	1	1	1	1	0	1	1	1	1	
22	20	0	1	0	1	1	0	1	0	0	0	
23	21	0	1	0	0	0	0	1	1	1	1	
24	22	0	0	0	0	0	0	0	0	1	1	
25	23	0	1	1	1	1	0	1	1	1	0	
26	24	1	1	1	1	1	1	1	1	0	1	
27	25	0	1	1	1	1	0	1	1	0	0	

Tal com indica l'enunciat, tots aquells subjectes amb una puntuació superior a 7 seran classificats com a subjectes amb alt rendiment, aquells que encerten menys de 3 ítems seran classificats com a subjectes de baix rendiment i la resta de subjectes formaran el grup de rendiment mitjà.

Per fer aquesta classificació podem utilitzar l'eina de filtre que hem comentat en l'apartat anterior.

1) Copieu els valors de la columna «Total» en la columna adjacent i canvieu el nom de la variable pel de «Grups».

2) Activeu l'opció de **Filtre** en el bloc de **Dades**.

3) Seleccioneu les puntuacions superiors a 7.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1													
2													
3	1	1	1	1	1	1	0	1	1	1	1	9	9
6	4	1	1	1	1	1	1	1	1	0	0	8	8
11	9	0	1	1	1	1	1	1	1	0	0	8	8
18	16	1	1	1	1	1	1	1	1	0	1	9	9
21	19	0	1	1	1	1	0	1	1	1	1	8	8
26	24	1	1	1	1	1	1	1	1	0	1	9	9

4) Assigneu la categoria «Alt rendiment» als subjectes seleccionats (podeu utilitzar l'opció d'autoemplenar).

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Ítems												
2		1	2	3	4	5	6	7	8	9	10	Total	Grups
3	1	1	1	1	1	1	0	1	1	1	1	9	Alt rendiment
6	4	1	1	1	1	1	1	1	1	0	0	8	Alt rendiment
11	9	0	1	1	1	1	1	1	1	1	0	8	Alt rendiment
18	16	1	1	1	1	1	1	1	1	0	1	9	Alt rendiment
21	19	0	1	1	1	1	0	1	1	1	1	8	Alt rendiment
26	24	1	1	1	1	1	1	1	1	0	1	9	Alt rendiment

5) Seleccioneu utilitzant l'opció de **Filtre** del bloc de **Dades** les puntuacions inferiors a 3.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Ítems												
2		1	2	3	4	5	6	7	8	9	10	Total	Grups
5	3	0	1	0	0	0	0	1	0	0	0	2	2
8	6	0	0	0	0	0	0	0	0	0	1	1	1
12	10	0	1	0	0	0	0	0	0	0	1	2	2
13	11	0	0	0	0	0	0	1	0	0	0	1	1
17	15	0	1	0	0	0	0	1	0	0	0	2	2
24	22	0	0	0	0	0	0	0	0	1	1	2	2

6) Assigneu la categoria «Baix rendiment» als subjectes seleccionats (podeu utilitzar l'opció d'autoemplenar).

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Ítems												
2		1	2	3	4	5	6	7	8	9	10	Total	Grups
5	3	0	1	0	0	0	0	1	0	0	0	2	Baix rendiment
8	6	0	0	0	0	0	0	0	0	0	1	1	Baix rendiment
12	10	0	1	0	0	0	0	0	0	0	1	2	Baix rendiment
13	11	0	0	0	0	0	0	1	0	0	0	1	Baix rendiment
17	15	0	1	0	0	0	0	1	0	0	0	2	Baix rendiment
24	22	0	0	0	0	0	0	0	0	1	1	2	Baix rendiment

7) Seleccioneu utilitzant l'opció de **Filtre** del bloc de **Dades** la resta de puntuacions (aquelles entre 3 i 7 punts).

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Items											
2		1	2	3	4	5	6	7	8	9	10	Total	Grups
4	2	0	1	0	1	1	1	1	1	0	0	6	6
7	5	0	0	0	0	1	0	1	0	0	1	3	3
9	7	0	1	0	1	1	0	1	1	0	1	6	6
10	8	0	1	0	1	1	0	1	0	0	1	5	5
14	12	1	1	1	1	1	1	1	0	0	0	7	7
15	13	0	1	0	0	1	0	1	1	0	0	4	4
16	14	0	1	1	1	1	0	1	1	0	1	7	7
19	17	0	1	0	1	1	0	1	1	1	1	7	7
20	18	0	0	0	0	1	0	1	0	1	0	3	3
22	20	0	1	0	1	1	0	1	0	0	0	4	4
23	21	0	1	0	0	0	0	1	1	1	1	5	5
25	23	0	1	1	1	1	0	1	1	1	0	7	7
27	25	0	1	1	1	1	0	1	1	0	0	6	6

8) Assigneu la categoria «Rendiment mitjà» als subjectes seleccionats (podeu utilitzar l'opció d'autoemplenar).

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Items											
2		1	2	3	4	5	6	7	8	9	10	Total	Grups
4	2	0	1	0	1	1	1	1	1	0	0	6	Rendiment mitjà
7	5	0	0	0	0	1	0	1	0	0	1	3	Rendiment mitjà
9	7	0	1	0	1	1	0	1	1	0	1	6	Rendiment mitjà
10	8	0	1	0	1	1	0	1	0	0	1	5	Rendiment mitjà
14	12	1	1	1	1	1	1	1	0	0	0	7	Rendiment mitjà
15	13	0	1	0	0	1	0	1	1	0	0	4	Rendiment mitjà
16	14	0	1	1	1	1	0	1	1	0	1	7	Rendiment mitjà
19	17	0	1	0	1	1	0	1	1	1	1	7	Rendiment mitjà
20	18	0	0	0	0	1	0	1	0	1	0	3	Rendiment mitjà
22	20	0	1	0	1	1	0	1	0	0	0	4	Rendiment mitjà
23	21	0	1	0	0	0	0	1	1	1	1	5	Rendiment mitjà
25	23	0	1	1	1	1	0	1	1	1	0	7	Rendiment mitjà
27	25	0	1	1	1	1	0	1	1	0	0	6	Rendiment mitjà

9) A partir de l'opció d'autoemplenar, calculeu el sumatori per la resta dels vint-i-cinc subjectes.

10) Un cop tingueu classificats els subjectes, copieu en un altre full les variables «ítem 1» (en lletres) i «Grups».

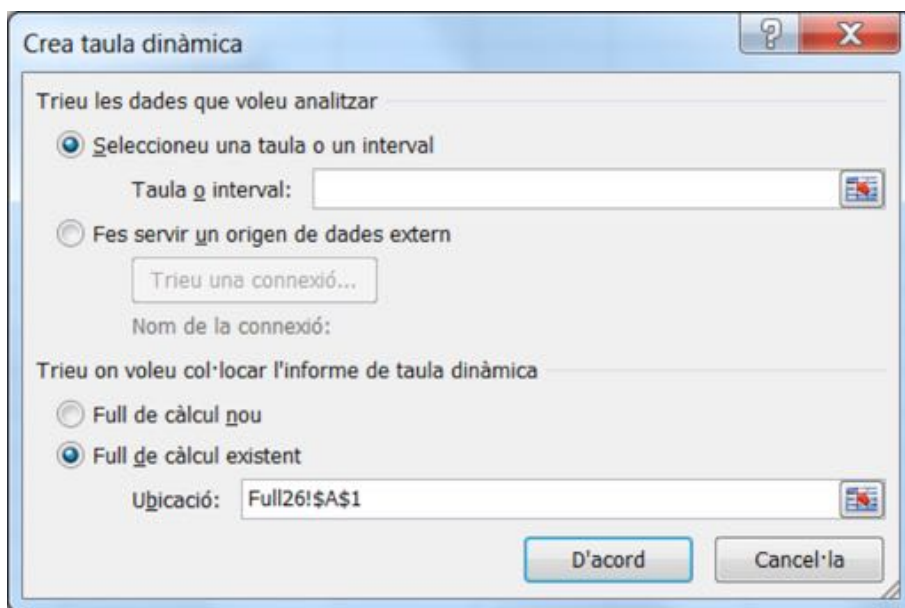
	A	B	C
1	Subjecte	ítem 1	Grups
2	1	A	Alt rendiment
3	2	B	Rendiment mitjà
4	3	C	Baix rendiment
5	4	A	Alt rendiment
6	5	D	Rendiment mitjà
7	6	D	Baix rendiment
8	7	B	Rendiment mitjà
9	8	C	Rendiment mitjà
10	9	C	Alt rendiment
11	10	B	Baix rendiment
12	11	B	Baix rendiment
13	12	A	Rendiment mitjà
14	13	C	Rendiment mitjà
15	14	D	Rendiment mitjà
16	15	B	Baix rendiment
17	16	A	Alt rendiment
18	17	D	Rendiment mitjà
19	18	D	Rendiment mitjà
20	19	B	Alt rendiment
21	20	B	Rendiment mitjà
22	21	C	Rendiment mitjà
23	22	C	Baix rendiment
24	23	C	Rendiment mitjà
25	24	A	Alt rendiment
26	25	B	Rendiment mitjà

A partir d'aquestes dades, caldrà construir una taula de contingència, mitjançant l'opció de **Taula dinàmica**.

Per obtenir-la seguiu la seqüència: **Inserció, Taula dinàmica, Taula dinàmica**.



Us apareixerà el quadre de diàleg següent:

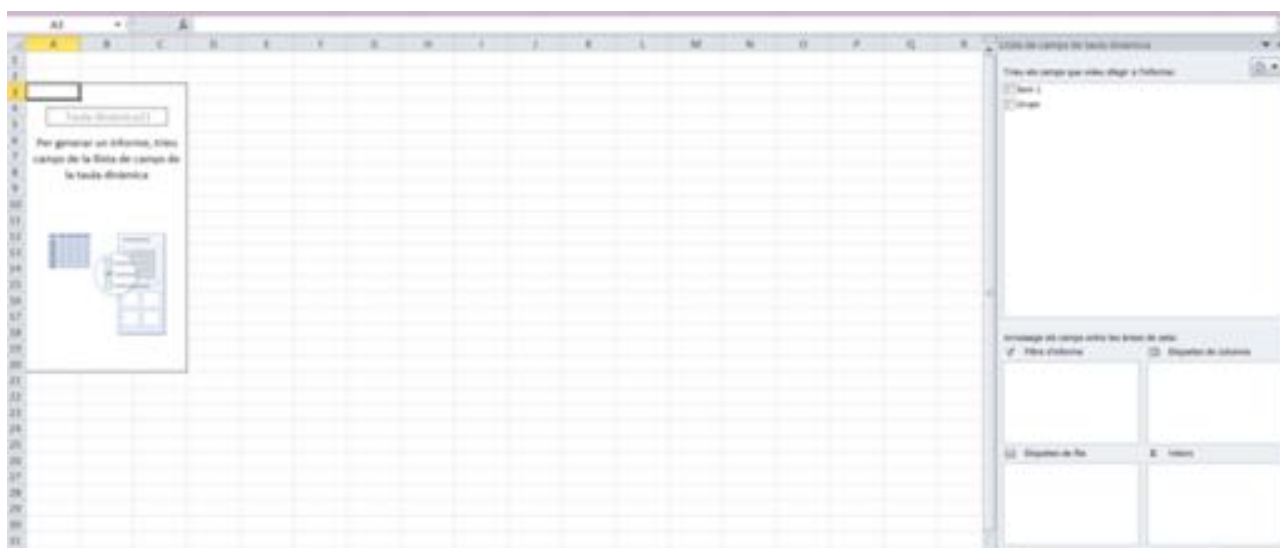


11) En el camp **Seleccioneu una taula o un interval** seleccioneu les caselles on tingueu les dades anteriors.

12) Per col·locar la taula, marqueu **Full de càlcul nou**, així us apareixerà en un nou full de l'Excel.

13) Premeu **D'acord**.

A continuació s'obrirà un nou full de l'Excel com la següent:



14) A la part esquerra teniu la taula per construir.

15) A la part dreta teniu els camps per omplir:

- a) Premeu sobre la variable «Grups» i arrossegueu-la fins a l'espai **Etiquetes de fila**.
- b) Premeu sobre la variable «Ítem 1» i arrossegueu-la fins a l'espai **Etiquetes de columna**.
- c) Premeu de nou sobre la variable «Ítems» i arrossegueu-la fins a l'espai **Σ Valors**.
- d) La taula de contingència us quedarà tal com apareix a la pantalla següent (les caselles de la taula de contingència que apareixen buides significa que el valor de la casella és 0).

Recompte d'item 1	Etiquetes de fila	A	B	C	D	Total general
Alt rendiment		4	1	1		6
Baix rendiment			3	2	1	6
Rendiment mitjà		1	4	4	4	13
Total general		5	8	7	5	25

Un cop heu obtingut la taula de contingència podeu utilitzar l'Excel com si fos una calculadora per obtenir l'índex de dificultat i de discriminació.

Per l'índex de dificultat:

- 1) Caldrà sumar els errors a partir de la funció de suma (recordeu que l'opció correcta és l'A).

Recompte d'item 1	Etiquetes de fila	A	B	C	D	Total general
Alt rendiment		4	1	1		6
Baix rendiment			3	2	1	6
Rendiment mitjà		1	4	4	4	13
Total general		5	8	7	5	25

=SUMA(C8:E8)

2) Dividiu el resultat per $K-1$ (nombre d'alternatives menys 1).

3) Resteu els encerts (total de l'opció A) del resultat anterior i dividiu pel total de subjectes que han contestat el test:

$$ID = \frac{A - \frac{E}{K-1}}{N} = \frac{5 - \frac{20}{3}}{25} = -0,07$$

Per l'índex de discriminació caldrà calcular la proporció de subjectes que encerten l'ítem en el grup d'alt i baix rendiment i calcular la diferència:

$$D = Pa - Pb = \frac{4}{6} - \frac{0}{6} = 0,67$$

En funció d'aquests valors, podem considerar que l'ítem 1 té una dificultat molt elevada i una alta capacitat de discriminació.

Per calcular els índexs de dificultat i de discriminació dels ítems 2 i 3 caldrà que seguim els mateixos passos descrits. A continuació, es mostra la taula de contingència i el resultat final d'aquests dos indicadors per cada un dels ítems.

Taula de contingència de l'ítem 2

Recompte d'ítem 2	Etiquet				
Etiquetes de fila	A	B	C	D	Total general
Alt rendiment			6		6
Baix rendiment		1	3	2	6
Rendiment mitjà	2		11		13
Total general	2	1	20	2	25

$$ID = \frac{A - \frac{E}{K-1}}{N} = \frac{20 - \frac{5}{3}}{25} = 0,73$$

$$D = Pa - Pb = \frac{6}{6} - \frac{3}{6} = 0,50$$

En aquest cas podem interpretar que l'ítem 2 és un ítem fàcil amb una discriminació alta.

Taula de contingència de l'ítem 3

Recompte d'ítem 3	Etiquet				
Etiquetes de fila	A	B	C	D	Total general
Alt rendiment				6	6
Baix rendiment	2	1	3		6
Rendiment mitjà	6		3	4	13
Total general	8	1	6	10	25

$$ID = \frac{A - \frac{E}{K-1}}{N} = \frac{10 - \frac{15}{3}}{25} = 0,20$$

$$D = P_a - P_b = \frac{6}{6} - \frac{0}{6} = 1$$

Podem interpretar que l'ítem 3 es tracta d'un ítem difícil amb una discriminació màxima.

5.2. Exercici 2

Seguint amb les consideracions fetes en l'exercici anterior, feu l'anàlisi dels distractors dels tres primers ítems.

Per a fer l'anàlisi dels distractors de cada un dels tres primers ítems, podem utilitzar les taules de contingència que hem confeccionat en l'exercici anterior i calcular els índexs de discriminació de cada una de les alternatives de respostes incorrectes utilitzant l'Excel com si fos una calculadora.

Anàlisi de distractors de l'ítem 1

1) Índex de discriminació de l'alternativa B:

$$D = P_a - P_b = \frac{1}{6} - \frac{3}{6} = -0,33$$

2) Índex de discriminació de l'alternativa C:

$$D = P_a - P_b = \frac{1}{6} - \frac{2}{6} = -0,17$$

3) Índex de discriminació de l'alternativa D:

$$D = P_a - P_b = \frac{0}{6} - \frac{1}{6} = -0,17$$

Els tres distractors tenen índexs de discriminació negatius, la qual cosa s'interpreta com que les tres alternatives de resposta presenten una discriminació adequada com a distractors. També cal comentar que l'ítem 1 és un ítem força difícil, i en conseqüència hi ha una proporció elevada de subjectes que escullen respostes incorrectes. Si observem aquestes respostes incorrectes, també podem concloure que les tres alternatives són adequades, atès que no hi ha grans diferències quant al nombre de subjectes que escullen cada alternativa. Així, l'alternativa B és escollida per vuit dels vint-i-cinc subjectes, la C per set, i la D per cinc.

Anàlisi de distractors de l'ítem 2

1) Índex de discriminació de l'alternativa A:

$$D = Pa - Pb = \frac{0}{6} - \frac{6}{6} = 0$$

2) Índex de discriminació de l'alternativa B:

$$D = Pa - Pb = \frac{0}{6} - \frac{1}{6} = -0,17$$

3) Índex de discriminació de l'alternativa D:

$$D = Pa - Pb = \frac{0}{6} - \frac{2}{6} = -0,33$$

En aquest cas, dos dels índexs de discriminació són negatius, i l'altre és igual a 0. Podem interpretar que les alternatives B i D presenten una discriminació adequada com a distractors, mentre que l'alternativa A no és capaç de discriminar entre els dos grups. De totes maneres, cal considerar que en aquest cas l'ítem té una dificultat molt baixa i, per tant, les alternatives incorrectes són escollides per molt pocs subjectes.

Anàlisi de distractors de l'ítem 3

1) Índex de discriminació de l'alternativa A:

$$D = Pa - Pb = \frac{0}{6} - \frac{2}{6} = -0,33$$

2) Índex de discriminació de l'alternativa B:

$$D = Pa - Pb = \frac{0}{6} - \frac{1}{6} = -0,17$$

3) Índex de discriminació de l'alternativa D:

$$D = Pa - Pb = \frac{0}{6} - \frac{3}{6} = -0,50$$

Els tres distractors presenten índexs de discriminació adequats. Malgrat tot, caldria considerar que l'alternativa B s'hauria de revisar, atès que només és escollida per un subjecte, la qual cosa ens indica que és massa evident que no és l'alternativa correcta.

Annex. Taules de distribució

PID_00204741

Julio Meneses

**Julio Meneses**

Professor dels Estudis de Psicologia i Ciències de l'Educació de la Universitat Oberta de Catalunya (UOC) i investigador de l'Internet Interdisciplinary Institute (IN3). Llicenciat en Psicologia per la Universitat d'Oviedo, es va doctorar en el programa de Societat de la Informació i el Coneixement de la UOC i, actualment, treballa com a responsable de la metodologia en diversos projectes d'investigació a les ciències socials. La seva recerca s'orienta a l'aplicació dels mètodes de recerca quantitativa i a l'ús de tècniques d'anàlisi multivariant, particularment en les àrees de la introducció de les TIC als sistemes educatius, la relació dels infants i els joves amb la tecnologia i l'abandonament i el reingrés dels estudiants universitaris.

Índex

1. Taules de distribució.....	5
--------------------------------------	----------

1. Taules de distribució

Taula 1. Distribució normal per a $x = x_1 + x_2$

Taula 2. Distribució χ^2 de Pearson

Taula 3. Distribució t de Student

Taula 4. Distribució F de Snedecor per a $\alpha = 0.900$

Taula 5. Distribució F de Snedecor per a $\alpha = 0.950$

Taula 6. Distribució F de Snedecor per a $\alpha = 0.975$

Taula 7. Distribució F de Snedecor per a $\alpha = 0.990$

Taula 8. Distribució F de Snedecor per a $\alpha = 0.995$

